E-kursuse "Eesti keele korpuste praktika" materjalid

Tartu Ülikoolis õpetatava aine FLEE02.100 "Eesti keele korpuste praktika" juurde

Aine maht 3EAP

Liina Lindström (Tartu Ülikool), 2009

Sisukord

1. Korpused ja korpuslingvistika. Eesti keele korpused.

Esimene teema on pühendatud korpustele ja korpuslingvistikale üldiselt. Lugege Kadri Muischneki koostatud lühike ülevaade korpuste kohta ning selle täienduseks John Sinclairi ülevaadet "**Corpus and Text: Basic Principles"** e-raamatust "Developing Linguistic Corpora: a Guide to Good Practice". (Toimetanud Martin Wynne)
http://ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

Korpused ja korpuslingvistika
Kadri Muischnek

Mis on korpus?
Keeleteaduseesmärk on loomuliku keele kirjeldamine. Ühed lingvistikakoolkonnad peavad selle all silmas eelkõige keelepädevuse (*competence*) ja teised hoopis keelekasutuse (*performance*) kirjeldamist. Nendel keeleuurijatel, kes arvavad keele olemuse avalduvat kasutuses, on uurimise aluseks tavaliselt mingi keelekogu – sõnaloend, näitelausete kogu jms. Koos arvutite muutumisega meie igapäevaelu osaks, on muutunud elektrooniliseks ka sellised keelekogud.
Keeleteaduses on sõna **korpus** all enne arvutite kasutuselevõttu tavaliselt mõeldud keeleainese kogumikku, mida kasutatakse uurimistöös materjalina (esineb see siis kartoteegi, lindikogu vms kujul) vastandina autori enda intuitsioonil põhinevatele üldistustele. Näiteks sõnaraamatut koostades või grammatikakirjeldust kirjutades ei piirdutud tavaliselt ainult oma keelepädevuse kirjeldamisega, vaid analüüsiti ka (tavaliselt ilukirjandusest) otsitud uuritavat sõna/nähtust sisaldavaid lauseid.
Arvutiajastul on korpusena hakatud mõistma peamiselt polüfunktsionaalseid elektroonilisel kujul olevaid tekstikogusid, [millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi kogu keelest].
Tekst ei tähenda siin ja edaspidi mitte ainult kirjalikku keelt, korpuses võib talletada ka suulist kõnet transkribeeritud kujul.
Paksus kirjas lõigus sisaldub tegelikult kaks korpuse definitsiooni: rangem ja realistlikum. Rangem definitsioon sisaldab ka nurksulgudes oleva teksti, st korpus on polüfunktsionaalne elektroonilisel kujul olev tekstikogu, millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi kogu keelest. Sellise korpuse koostamiseks tuleb kõigepealt mingi ajavahemiku keelekasutus jagada tekstiklassidesse ning määrata kindlaks iga tekstiklassi osakaal ning koostada korpus selliselt, et tekstiklassid oleksid korpuses esindatud vastavalt nende osakaalule kogu keelekasutuses. Selline korpus on **representatiivne** e esinduslik valitud ajavahemiku keelekasutuse suhtes. Selliselt koostatud korpust nimetatakse ka **suletud korpuseks**, sest kui selline korpus on kord valmis, tehtud, ei saa sinna enam tekste juurde lisada või neid korpusest ära võtta ilma, et kaoks tekstiklasside vaheline tasakaal.
Et saavutada suletud korpuse representatiivsust, on vaja:
1) defineerida, mida st millist keelt peab see korpus esindama
2) liigitada see keel mingite tunnuste alusel tekstiklassidesse
3) määrata kindlaks iga tekstiklassi hulk ja/või mõju meid huvitaval perioodil

4) selle järgi määrata kindlaks selle tekstiklassi osakaal korpuses
Representatiivsus ei saa aga olla absoluutne, st korpus ei saa olla esinduslik **kogu** ajaperioodi keelekasutuse suhtes. Esiteks jaguneb keelekasutus suuliseks ja kirjalikuks ning nö tavainimene räägib/kuuleb keelt rohkem kui kirjutab/loeb, st domineerib suuline keelekasutus. Kuid suulise keele korpuse koostamine on palju töömahukam ja seega ka kallim kui kirjaliku keele korpuse koostamine. Teiseks on lisaks üldkeelele olemas ka erialakeeled oma sõnavara ja muude eripäradega. Kuigi suulise keele korpused muutuvad aina suuremaks, vaadatakse representatiivse korpuse koostamisel suulise-kirjaliku keelekasutuse proportsioonidele praktikas ikka läbi sõrmede. Tavaline praktika on koostada eraldi suulise ja kirjaliku keele korpused.
Keel, eriti sõnavara on aga muutuv ja teatud ajavahemiku tekste sisaldav suletud korpus pole 10 aastat hilisema keele kohta enam representatiivne.
Vabama tõlgenduse kohaselt võib korpuseks nimetada ka lihtsalt mingit kogumit tekste elektroonilisel kujul kindlas elektroonilises formaadis. Sel juhul ei ole tekste valitud kindlaid põhimõtteid või eesmärke silmas pidades, vaid neid on kogutud selleks, et kasutaja võiks talletatud tekstide hulgast teha valikuid vastavalt oma vajadustele või on lihtsalt kogutud seda, mida on olnud võimalik/lihtne koguda. Sellisesse **avatud** korpusesse, erinevalt suletud korpusest, saab tekste pidevalt juurde lisada.
Kasutaja seisukohalt on suletud korpus „rohkem valmis“ kui avatud korpus: suletud korpus esindab mingi perioodi ja/või allkeele keelekasutust ja selle representatiivsuse eest on korpuse koostajad juba hoolitsenud. Avatud korpuse kasutaja peab vajaduse korral endale representatiivse tekstikogumi pakutavatest allkorpustest ise kokku komplekteerima.

Korpuste liigid

Keelekorpusi saab liigitada mitme tunnuse alusel. **Suletud** vs **avatud** korpusest oli juba juttu, **suulise** vs **kirjaliku** keele korpustest samuti. Prototüüpse suulise (st spontaanse) ja prototüüpse kirjaliku (st planeeritud) keelekasutuse vahele on viimasel aastakümnel tekkinud uus tekstiliik – kirjalik kuid spontaansete sugemetega nn **uue meedia** keelekasutus (jututoad, foorumid, kommentaarid jms).
Keeleteaduse ja arvutilingvistika eesmärkidest lähtuvalt koostatakse mitmesuguseid **erikorpusi** – mingit kindlat allkeelt esindavaid või spetsiaalselt märgendatud tekstikogusid. Näiteks vajavad murdeuurijad murdekorpust, vana kirjakeele uurijad vana kirjakeele korpust, keele arengut pikema ajavahemiku jooksul võimaldab jälgida diakrooniline korpus, jne.
Paralleelkorpus sisaldab teksti ja selle tõlget (tõlkeid), kusjuures tavaliselt on paralleelkorpus joondatud ehk paralleelistatud, st on näidatud, milline lause(osa) on millise lause(osa) tõlkeks. Paralleelkorpusi kasutatakse võrdlevas keeleuurimises, nad on heaks abivahendiks sõnaraamatute koostamisel, masintõlke „kuum“ suund statistiline masintõlge vajab väga suuri (mitukümmend kuni mitusada miljonit sõna) paralleelkorpusi.
Erialakeelte uurijad kasutavad erialakeelte korpusi, erialakeele mitmekeelse sõnaraamatu koostamisel on hea aluseks võtta erialakeele paralleelkorpus.

Korpuste märgendamine

Märgendamiseks nimetatakse eksplitsiitse info lisamist korpusesse. Tavalisim märgendus on lausepiiride tähistamine, selle tulemusel saab nt korpuse kasutajaliidese kaudu päringule vastuseks terviklause. Eesti keele puhul on oluline morfoloogiline

märgendamine, mille käigus lisatakse igale tekstisõnale tema lemma ja info grammatiliste kategooriate kohta (sõnaliik, kääne, pööre, arv jms). Suurte korpuste käsitsi märgendamine nõuaks liiga palju inimtööd ja seda püütakse automatiseerida. Näiteks saab lausepiire ja morfoloogilist infot eestikeelsesse (kirjakeelsesse) teksti lisada täisautomaatselt vastavate programmide abil. Muidugi pole ükski programm täiuslik, näiteks eesti keele automaatsel morfoloogilisel märgendamisel saab umbes 5% tekstisõnadest vale analüüsi. Keerulisemaid või vähemuuritud lingvistilisi nähtusi tuleb siiski märgendada käsitsi või poolkäsitsi.

Korpuse suurus

sõltub väga paljudest asjaoludest. Arvutilingvistika vajab üldjuhul suuremaid korpusi kui lingvistika, arvutilingvistide hulgas liigub selle kohta raskesti tõlgitav lendlause *There is no data like more data*. Kirjaliku üldkeele korpused on tänapäeval tüüpiliselt väga suured avatud korpused, hetkel suurim paistab olevat sakslaste korpuste kogum *Mannheimer korpora*, milles on 2009. aasta alguse seisuga 3,6 miljardit sõna. „Korraliku“ tänapäevase korpuse suurus algab sajast miljonist sõnast. Tänapäeva kirjaliku eesti keele korpuses (Koondkorpus) on umbes 250 miljonit sõna. Suulise keele korpused on reeglina palju väiksemad, ühe miljoni sõna suurune suulise keele korpus on juba vägagi arvestatav keeleressurss. Muidugi on palju väiksemad käsitsi märgendatud erikorpused.

Mis on korpuslingvistika?

Terminil on kaks sisu: korpuspõhine lingvistika ja korpuste koostamist käsitlev distsipliin.

Korpuslingvistika kui korpuspõhine lingvistika on keeleteaduse suund, mis rõhutab tegelikust keelekasutusest pärineva materjali kasutamise tähtsust keeleuurimisel. Muidugi ei ole korpuslingvistika omaette lingvistika haru selles mõttes, nagu seda on näiteks kognitiivne lingvistika või konstruktsioonigrammatika, st ta ei ole keeleteooria vaid keele uurimise viis.

Korpuslingvistika kui arvutilingvistika haru tegeleb arvutikorpuste koostamispõhimõtete, ja –tehnoloogiatega, korpuste märgendamise ning automaatse analüüsi vahendite väljatöötamisega.

2. Kirjakeele korpuse kasutajaliidese kasutamine: otsing märgendamata tekstist märgijada põhjal.

Korpuse kohta võib lugeda 1. teemas toodud artiklitest ja lehelt http://www.cl.ut.ee/korpused/. Kasutajaliidese kaudu saab kasutada enamikku neist:

Eesti kirjakeele korpus 1890-1990, nn monitorkorpus; sisaldab ajakirjandus- ja ilukirjandustekste (u 2000 sõna igast ilukirjandusväljaandest vaadeldaval kümnendil ning mõningad valitud ajalehed tervikkujul vaadeldaval kümnendil)

Tasakaalus korpus (ajakirjandus+ilukirjandus+teadus) - sisaldab u 5 miljonit sõna igast valdkonnast

Eesti keele koondkorpus - sisaldab kõikvõimalikke tekste, eesmärk on tekstimassiivi võimalikult suureks ajada.

Loe kõigepealt läbi kasutusjuhend ning tutvu kasutajaliidesega http://www.cl.ut.ee/korpused/kasutajaliides/. Erineva info kättesaamiseks korpustest tuleb tihtipeale otsida mitte lihtsalt tekstisõna järgi, vaid mõelda välja natuke üldisem päring. Teema materjal puudutab erisümbolite jms kasutamist päringute tegemisel.

Käesoleva kursuse raames tegeleme eelkõige taoliste otsingutega, mis on lähtematerjaliks edaspidiseks lingvistiliseks uurimistööks. Seetõttu on ka ülesandepüstitus üldjuhul mingi keelelise nähtuse keskne.

NB! Iga päringu tegemisel tuleb arvestada kahe asjaoluga:

1) päring peab maksimaalselt hästi leidma üles kõik teie uurimisülesande seisukohalt vajalikud kasutusjuhud ega tohi teha mingit süstemaatilist viga (näiteks mingit konkstruktsiooni otsides välja jätta teistsuguse sõnajärjega lauseid vms);

2) päringus tuleb minimeerida "prahi" hulk, et mitte teha liiga palju käsitsitööd. Siiski alati ei ole võimalik saada 100% head või täpset tulemust; sel juhul pigem teha pisut rohkem käsitsitööd kui süstemaatiliselt jätta välja selliseid kasutusjuhtumeid, mis tegelikult teie uurimistöö seisukohalt on olulised.

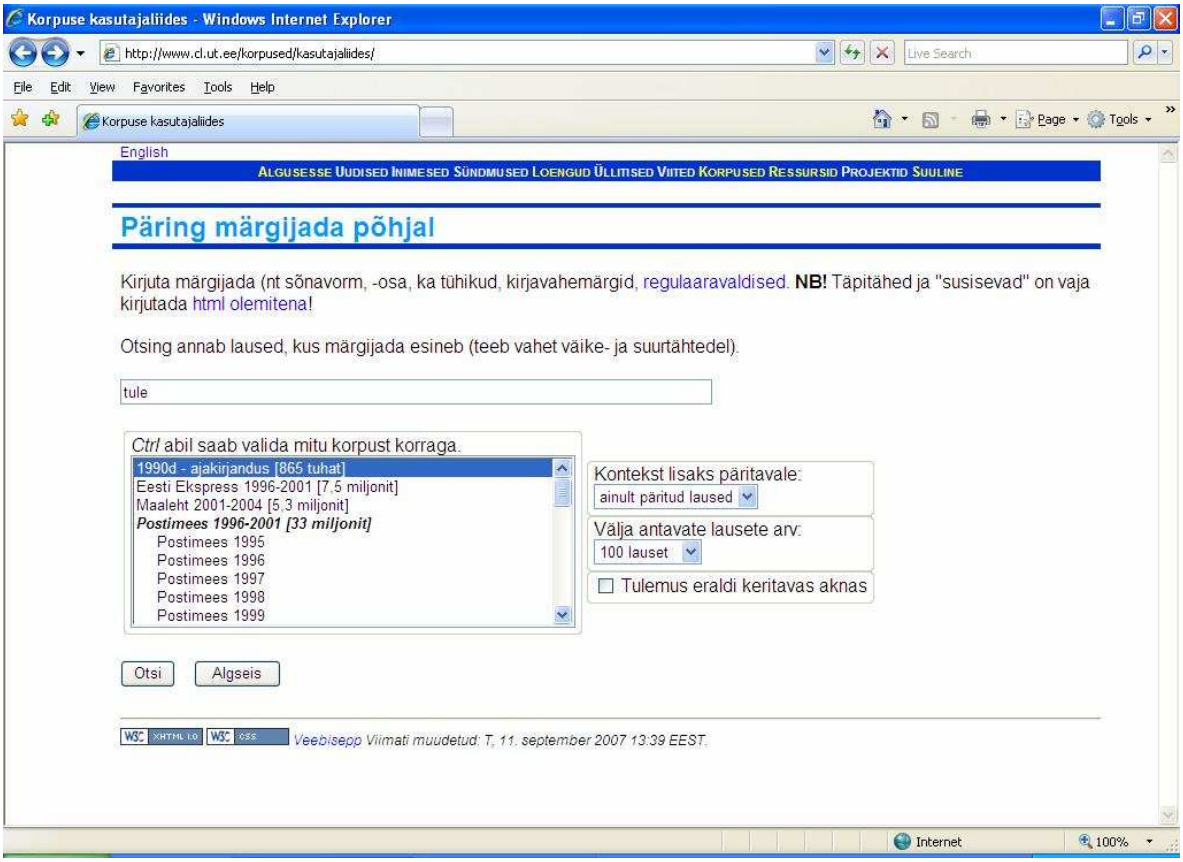
Kirjakeele korpuse kasutajaliidese kasutamine

- Ava internetis TÜ arvutilingvistika uurimisrühma lehekülg **http://www.cl.ut.ee**
- vali Korpused, tutvu, millistest osadest kirjakeele korpus koosneb
- vali ülalt menüüst Ressursid --> Eesti kirjakeele korpuse kasutajaliides

Kasutajaliides võimaldab otsida tekstidest **märgijada** põhjal, st otsitakse mingi tekstilõigu põhjal. Otsimiseks tuleb valida sobiv (alam)korpus, mitmest alamkorpusest otsimiseks tuleb valimise ajal Ctrl-klahvi all hoida. Rippmenüüdest saab vajadusel lisada kuni 5 lauset kontekstiks, see aitab paremini mõista lause viitsuhteid vms. Rippmenüüst saab valida ka, mitu lauset väljundiks väljastatakse. NB! Harjutamise eesmärgil on mõistlik väljundi hulka piirata! Kui väljundiks hulk on piiratud, ei anta väljundiks mitte järjestikuseid lauseid, vaid need nopitakse välja kogu korpusest, nt iga 36. lause.

Päringu tulemuseks saadakse read, mis sisaldavad otsitud lõiku. Iga rea alguses on kood, mis ütleb, missugusest tekstist antud rida pärit on. Tulemuste lehekülje alguses on raport, milline oli päring, kui paljule päringule vastavaid ridu leiti ning iga mitmes rida kuvatakse ekraanile.

Näide:



otsing 'tule' annab tulemuseks ridu (=lauseid), milles on märgijada *tule*:

Otsiti	Leiti
Päritud märgijada tule **Päritud allkorpus** 1990_ajalehed_26_08_04 **Kontekstilauseid ees ja järel** pole tellitud	**Kokku ridu päritud failis:** 71335 **neist päringule vastas:** 3548 **Väljastatakse lauseid:** kuni 100 **Näidatakse iga 36. lause**

AJAE1980\tat0516 Osa inimesi on nagu pilvedes, ei näe enam oma tööesist, oma igapäevast tööd, mida ka ju ikka paremini korraldada ja jõukohasemaks sättida **tule**b.

AJAE1980\tat0516 Mida ma oma **tule**viku vallalt sooviksin?

AJAE1980\tat0516 Mis tegelikult **tule**b?

AJAE1980\tat0516 Nüüd on jälle nii otsustatud, et ehitatakse siiagi ühepereelamuid, eks siis hakka inimesi juurde **tule**ma.

AJAE1980\tat0516 Oma saadikult nõuaksin: Puhjasse **tule**b ruttu ehitada tööstuskaupade kauplus.

AJAE1980\tat0516 Ja see on ka üks suur rumalus, et meie haigla ära kaotati, **tule**b tagasi taotleda.

AJAE1980\tat0516 Buss **tule**b ja viib nad Lähtele.

AJAE1980\tat0516 Eks andmed **tule** külanõukokku, avalikuks pole veel midagi saanud.

AJAE1980\tat0516 Oma inimestele **tule**b maal korralikud elutingimused luua ja õiget palka maksta, siis pole võõraid tarvis.

AJAE1980\tat0516 Midagi tervele rajoonile vajalikku ei **tule** praegu meelde.

AJAE1980\tat0517 Tema nimega seostatakse enamikku **tule**vikulootustest, temalt ootavad lunastust ka Saksa DV elanikud.

AJAE1980\tat0517 Gorbatšov on nagu messias, kelle **tule**ku on inimkond jõudnud lõpuks ära oodata.

Oluline on meeles pidada, et **väikestel ja suurtel tähtedel tehakse vahet**. Kui otsitav lõik võib sisaldada nii väike- kui suurtähti (nt lause algul), tuleb arvestada mõlema võimalusega.
Märgijada järgi otsides tuleb arvestada ka **tühikute ja kirjavahemärkidega**. Kui eelnevat otsingut piirata nii, et *tule* ees on tühik ' tule', saame tulemuseks vaid need read, kus *tule* on sõna algul.
Kui otsitav lõik sisaldab **täpitähti**, tuleb need sisestada html-kujul, vastasel korral päringule vastavaid ridu ei leita. Täpitähtede tabel:
Š = Š
š = š
Ž = Ž
ž = ž
Õ = Õ
õ = õ
Ä = Ä
ä = ä
Ö = Ö
ö = ö
Ü = Ü
ü = ü

Otsingut saab modifitseerida ka üldisemaks, nii et üks sümbol asendaks mitut märki. Selliste regulaaravaldiste tegemiseks kasutatakse erisümboleid.

Erisümbolid

.	üks suvaline märk, nt päring 'tule. ' annab vastuseks read, kus on märgijada *tule* ja veel üks sümbol, millele omakorda järgneb tühik, nt *tulen* , *tuled* , *tuleb* , *tulek* , aga mitte *tulemine* vms.
.*	mistahes sümbol null kuni lõpmatu arv kordi, proovi: '.*'
x*	x esineb null kuni lõpmatu arv kordi, nt päring ' ku*' annab vastuseks nii *k*, *ku*, *kuu*, *kuuu*, *kuuuu* (kui selliseid sõnu korpuses oleks)
x+	otsib märgijadasid, milles x esineb vähemalt 1 korra (maksimum pole piiratud) nt päring ' ku+' annab vastuseks *ku*, *kuu*, *kuuu*, *kuuuu* jne
[ ]	kasutatakse märkide valiku näitamiseks, nt 'm[ae]' otsib märgijadasid, kus on kõrvuti kas *ma* või *me*.
[a-z]	inglise tähestiku väiketähed, nt päring ' [a-z][a-z] ' annab vastuseks ridu, milles on 2-tähelisi sõnu, milles pole täpitähti: *on*, *ja*, *et*, *no* jne
[A-Z]	inglise tähestiku suurtähed
[0-9]	kõik numbrid
[^a]	kõik märgid, välja arvatud *a*, nt päring '[^]ma' annab vastuseks ridu, milles ma paikneb kusagil sõna siis, mitte sõna alguses (s.t tema ees ei saa olla tühikut)
[^a-zA-Z]	mistahes sümbol, v.a tähestiku tähed
x{2,5}	otsib märgijadasid, kus x-i esineb minimaalselt 2, maksimaalselt 5 korda järjest, proovi nt 'ku{1,3}'
x{2}	otsib märgijadasid, milles x-i esineb vähemalt 2 korda järjest (maksimum pole piiratud), nt 'ku{2}'
\?	kaldkriips tühistab järgneva märgi erisümbolitähenduse. See on vajalik näiteks siis, kui otsida küsilauseid, st lauseid, mille lõpus on küsimärk: ' \?'
(ah)	sulgude vahele kirjutatakse märgijada, millele hiljem saab viidata, näiteks märkida selle jada kordust (seda märgib kaldkriipsuga arvsulgude järel): päring (ah)\1 annab vastuseks *ahah*
\b	leiab sõna alguse ja lõpu. Sõna defineeritakse kui tähtede-numbrite ja allkriipsude jada, nii et sõna lõppu tähistab tühik või märk, mis ei kuulu tähtede, numbrite ega allkriipsu hulka. Kuna täpitähed ja susisevad sisaldavad ampersandi (&) ja semikoolonit (;), siis ei kuulu nad selle definitsiooni kohaselt sõna hulka, nagu ka sidekriips. Nurksulgude vahel tähistab \b tagasiaste (backspace) klahvi.
|	võimaldab otsida korraga kaht märgijada, nt 'mees|naine' otsib ridu, kus leidub kas märgijada *mees* või *naine*. Mõlemal pool toru võib olla ka regulaaravaldis.

vt ka http://www.cl.ut.ee/korpused/kasutajaliides/erispikker#reg

Ülesanded koos vastustega

Tehke ülesanded kõigepealt iseseisvalt, seejärel kontrollige vastuseid ja vaadake kommentaare.

NB! Vastustes ○ tähistab tühikut!

1. Vaata, kuidas on tekst korpustesse sisestatud: milliseid jutumärke kasutatakse, kas kirjavahemärgid on eraldatud tekstist tühikuga või mitte (erinevates korpustes võivad need olla erinevalt).

Korpusetes on enamasti kirjavahemärgid muust tekstist tühikuga eraldatud.
Jutumärkidest on kasutatud nii tavalisi jutumärke " kui ka << tekst >>

2. Otsi ilukirjanduse korpustest otsest kõnet. NB! Erinevates korpustes on erinevad jutumärgid.

Otsest kõnet võib otsida jutumärkide abil, ent jutumärke kasutatakse ka mujal. Näiteks võib päringu vastuseks tulla palju kaupade, toodete, ettevõtete jne nimetusi, mida praegu ei pea enam kirjakeele reeglite järgi jutumärkide vahele kirjutama.
Seda arvestades on kõige kindlam otsida otsest kõnet saatelausele järgneva kooloni ja jutumärkide abil:
:○<<○.*○>>
: ○"○.*○"
Tuleb arvestada, et otsese kõne osa võib ületada ka lause piiri, s.t otsese kõne lõpumärke ei pruugi samas lauses olla.
:○<<
:○"

3. Otsi lauseid, milles on kasutatud araabia numbreid.
Araabia numbrite otsimiseks saab kasutada erisümbolit [0-9]. Konks on aga selles, et kui päringureale sisestada lihtsalt [0-9] , saame tulemuseks kõik read, sest iga rea/lause alguses on kood, milles on araabia numbreid kasutatud. Seega tuleb välja mõelda, kuidas koodist nö üle astuda.
Selleks tuleb teada, et koodile järgneb 4 tühikut. Me peame otsima järjendeid, milles oleks 4 tühikut, seejärel võib, aga ei pruugi veel olla midagi (sest araabia numbrid ei ole ju tingimata lause alguseses), ja siis on araabia number/numbrid. Sobiv päringurida oleks järgmine:
○○○○.*[0-9]
Seda on võimalik esitada ka kompaktsemalt, nii et tühikule looksulgudes järgnev arv näitab, mitu korda seda esineb:
○{4}.*[0-9]
Kui teid häirib, et lause alguses olev kood läheb rasvasesse kirja, võib päringut esitada nii:
(?<=○{4}).*[0-9]

See komplekspäring täpsustab, et otsitava lõigu EES on 4 tühikut, mitte otsitava lõigu ALGUSES on 4 tühikut.

4. Otsi lauseid, milles ei ole kasutatud araabia numbreid.
See ülesanne tundub analoogiline eelmisega, ent siin on siiski konks. Kui me teeksime samasuguse päringurea nagu enne, aga lisaksime märgi *välja arvatud* [^0-9], saaksime ikkagi päringu vastuseks ka need read, kus araabia numbreid on. Proovi:
○{4}.*[^0-9]
Põhjus peitub selles, et eelmises ülesandes kasutasime märki .*, mis põhimõtteliselt võib hõlmata ka araabia numbreid. Samuti ei otsinud me mitte rea lõpuni, vaid ainult araabia numbriteni. Me peame aga olema kindlad, et kuni rea lõpuni araabia numbreid pole. Selleks on sobiv kasutada rea lõpu märki $. Sobiv päringurida oleks seega
○{4}[^0-9]*$

5. Otsi ühe käsuga välja relatiiv-interrogatiivpronoomeniga *kes* (*kelle, keda*) ning *mis* (*mille, mida*) algavad kõrvallaused.
Kõigepealt tuleb arvestada, et mõlemad on käänduvad sõnad:
kes *mis*
kelle *mille*
keda *mida*
kellele *millele*
jne, ülejäänud käändevormid moodustatakse *kelle/mille*-tüvest.
Alustame kõigepealt kes-sõna vormide ühe otsinguga leidmisest, siis vaatame, kuidas need kokku panna.
Tuleb leida kõigi sõnavormide ühisosa, selleks on *ke*
Seejärel muutuv osa: 3. sümbol on kas s, l või d, võime päringu koostada järelikult nii:
ke[sld]
Ülesandes küsitakse kõrvallause alguses olevaid sõnu. Väheste eranditega kirjutatakse need kõrvallause algul kõik koma järele, seega saame päringu teha nii:
,○ke[sld]
Analoogiliselt tuleb toimida ka *mis*-sõnaga:
,○mi[sld]
Et neid korraga pärida, on kõige mõistlikum kasutada "toru", siis otsitatakse kas toru ees või toru järel olevat järjendit, aga mitte mõlemat korraga:
,○ke[sld]|,○mi[sld]
Nagu näha, on mõlemas veel ühisosa , ja tühik. Me võime päringut veelgi komplekssemaks muuta ning viia alternatiivsed osad sulgude vahele, ühisosa jääb väljaspoole sulge, näiteks nii:
,○(ke[sld]|mi[sld])
või nii:
,○(ke|mi)[sld]

6. Otsi kõrvutiasetsevaid täpitähti (sh suurtähed).
Ülesanne oleks väga lihtne, kui täpitähed poleks html-kujul. Aga nad on html-kujul, ja me peame seda arvestama.
õ = õ / Õ = Õ

ä = ä / Ä = Ä
ö = ö / Ö = Ö
ü = ü / Ü = Ü
Seega enam-vähem ühtmoodi kirjutatakse äöüÄÖÜ, teistmoodi õÕ. Päringus võibki seda arvestada.
Eesti keele kohta teame veel, et võimalikud ei ole järjendid õü, õä,öö – seega kui esimene on õ, on ka teine täht õ (nt *võõras*). Saame pärida nt eraldi täpitähtede ja õ-de kohta:
&[aouAOU]uml;&[aouAOU]uml;
&[oO]tilde;&[oO]tilde;
ja need siis kokku panna ühte päringusse:
&[aouAOU]uml;&[aouAOU]uml;| &[oO]tilde;&[oO]tilde;
või elegantsemalt:
(&[aouAOU]uml;|&[oO]tilde;)\1
Lõpus \1 tähendab, et eelnevat järjendit korratakse veel üks kord.

7. Otsi vähemalt kolme vokaali ühendeid (sh suurtähed, täpitähed jäta välja). Mitu vokaali võib korpuste põhjal maksimaalselt sõnas järjestikku olla?
Täpitähtedega oleks keeruline, sest need on html-kujul, seepärtast on need ülesandest välja jäetud.
Muude vokaaliühenditega on soovitav aga kasutada vokaalide loendit ja selle järele lisada, mitu korda loendis olevaid süboleid peaks kõrvuti olema. Kõige pikema vokaaliühendi leidmiseks võib seda numbrit muuta seni, kuni korpusest päringule veel vastust leiab:
[aeiouAEIOU]{3}

8. Otsi korpustest sõnu, mille pikkus on vähemalt 20 tähte (täpitähed võiks segaduste vältimiseks välja jätta). Mis on kõige pikem sõna?
Täpitähede väljajätmiseks arvestame ainult neid sõnu, milles on inglise tähestiku tähed. Selleks kasutame erisümbolit [a-zA-Z]:
[a-zA-Z]{20}

9. Otsi 1970ndate ilukirjandusest lauseid, milles on kasutatud jussiivi eitavat vormi (*ärgu tulgu, ärgu nähku*).
Jussiivi otsimiseks piisab tegelikult vaid ühe sõna otsimisest – vormi ärgu muudes kontekstides/funktsioonides ei kasutata. Seega lihtsaim otsing oleks
○&[Aa]uml;rgu○

10. Otsi progressiivkonstruktsiooni *olema* +*-mas* (*on valmimas*).
Siin tuleb kõigepealt mõelda, mis vormides saab *olema*-verb selles konstruktsioonis olla. Tuleks arvestada vähemalt kõiki kindla kõneviisi oleviku ja lihtmineviku vorme:
olen – olin
oled – olid
on – oli
oleme – olime
olete – olite
on – olid

Seejärel tuleks erisümbolite abil need vormid kuidagi kokku võtta, leida, mis neis on ühist:
○o[ln][ei]*
mas-tunnuse otsimine on suhteliselt lihtne, see lõppe ei saa varieeruda. Tähtis on vaid meeles pidada, et see paikneb sõna lõpus (järgneb tühika, aga tühik ei tohi olla *mas*-i ees, *mas*-i ees tohib olla minimaalselt 3 tähte nagu sõnas *olemas, söömas*).
Edasi tuleks mõelda, mis nende kahe vahel võib olla. Et sõeluda välja võimalikud just otsitava konstruktsiooni liikmed muudest *mas*-idest, peaksid olema-verb ja mas-vorm kindlasti paiknema samas osalauses. Osalausete piiri on raske määrata, ent üheks võimaluseks on kasutada selleks osalausete piiridel paiknevat koma. Otsime nii, et olema- ja mas-vormi vahel ei paikneks koma, küll võib seal olla kõike muud piiramata hulgal. Seega võiksime päringurea koostada nii:
○o[ln][ei]*[^,]*mas○
See otsing ei tööta loomulikult veatult. Konstrutksioone otsides tuleb üldse arvestada päris suure veaprotsendiga ning tegelikku materjali otsides tuleb pidevalt mõelda sellele, ega mõnda kasutusvõimalust ei ole süstemaatiliselt "maha magatud".
Praegusel juhul on näiteks süstemaatiliselt välja jäetud võimalus, et lauses on teistsugune sõnajärg: nt täiesti võimalik, et *olema*-verb ja *mas*-vorm on teises järjekorras (*olemas oli*). Päringu täiustamiseks võiks seega teises järjekorras variandi esitada alternatiivse otsinguna, ent sel juhul kaotab mõtte mas-vormi ees piirang [^,]*, mõttekam oleks piirata nii, et mas-i ees oleks vähemalt 3 mittetühikut: [^○]{3}
○o[ln][ei]*[^,]*mas○|[^○]{3}mas○[^,]*○o[ln][ei]*
Seda päringurida katsetades näeme aga, et liiga palju tuleb nö prügi sisse. Seega tasuks eelnevalt katsetada, kas mõlemat sõnajärjevarianti on vajalik arvestada ning mõelda ja otsustada, kas otsida välja rohkem ja pärast käsitsi sorteerida mittevajalik välja või otsida välja pigem vähem ja arvestada, et on võimalus, et osa variante ei ole leitud. Lähenemine sõltub peamiselt töö eesmärgist.

3. TÜ eesti kirjakeele korpuse morfoloogiliselt märgendatud alamkorpus. Keeleveeb.

Tegeleme kirjakeele morfoloogiliselt märgendatud tekstidega, mille jaoks on kasutajaliides aadressil http://www.cl.ut.ee/korpused/morfliides/.

Morfoloogiliselt märgendatud korpuse koostisosade jms vt ka http://www.cl.ut.ee/korpused/morfkorpus/

Morfoloogiliselt märgendatud tekstidest otsides saab lisaks märgijadale otsida ka morf. märgendite põhjal. Otsimiseks tuleb teada morf. lühendeid ja märgendamise põhimõtteid. Lühendite nimekirja leiad kasutajaliidese juurest http://www.cl.ut.ee/korpused/morfliides/seletus. Vaata lühendite nimekirja ja proovi selle järgi otsida. Tuleb arvestada, et selles kasutajaliideses ei tehta vahet, kas otsitav märgijada (sõna, sõnaosa vms) kuulub sõnasse või selle grammatilisse infosse, seega võib otsida mõlema info järgi. Ka siit otsides on kasulik teada erisümboleid, mille abil saab otsingut üldisemaks muuta.

Kursuse materjal sisaldab infot ka keskkonnas www.keeleveeb.ee paikneva morfoloogiliselt märgendatud teksti põhjal tehtava päringuvormi kohta.

Teema juurde kuuluvad ka ülesanded koos vastustega.

Morfoloogiliselt märgendatud tekstid

Morfoloogiliselt märgendatud korpus: http://www.cl.ut.ee/korpused/morfliides/

Morfoloogiline märgendamine tähendab seda, et tekstid on varustatud lisainfoga sõnaliigi ja muutevormi kohta.

Morfoloogiliselt märgendatud korpuses on tegemist tekstidega, mille on märgendanud automaatne morfanalüsaator, ning seejärel on tekstid ühestatud, st kõikidest sõna võimalikest analüüsivariantidest on valitud välja õige. Tekste on märgendanud kaks märgendajat, seejärel on kolmas märgendaja vaadanud üle erinevused, mis kahe märgendaja märgenduse vahel on olnud, ning otsustanud ühe variandi kasuks. Seega peaks tekstide märgendus olema usaldusväärne.

Näide : Olin ole+in //_V_ aux indic impf ps1 sg ps af // uurinud uuri + nud //_V_ main partic past ps // niisugust nii_sugune+t //_P_ sg part // pilti pilt+0 //_S_ com sg part // , , //_Z_ Com //

Märgendus koosneb järgmistest osadest (NB! ○ märgib tühikut):

sõna○**tüvi+lõpp**○//○**analüüs**○//

sõna = sõne (nagu see tekstis esineb)

tüvi (verbidel *ma*-infinitiivi tüvi, käändsõnadel nimetavas käändes tüvi), plussiga on eraldatud muutlõpp

// eristab morf infot muust (morf info alguses ja lõpus)

_V_ sõnaliik (antud juhul verb)

Näiteks sõna *elanud* (ühendis *on elanud*) on kirjeldatud nii:

elanud ela+nud //_V_ **main partic past ps** //

Sõnaliikide ja vormikirjelduste lühendeid vaata siit:
http://www.cl.ut.ee/korpused/morfliides/seletus

Ka selle korpuse puhul tuleb arvestada sellega, et täpitähed on html-kujul, samuti saab siin kasutada erisümboleid ja regulaaravaldisi, vt
http://www.cl.ut.ee/korpused/kasutajaliides/erispikker

Morf märgendatud korpusest otsides tuleb silmas pidada, et päringuauk ei saa ise aru, kas tahad otsida tekstisõna või morfoloogilise info osast, st ta käsitleb kõike kui üht märgijada. Kui on vaja otsida vaid morfoloogilise info seast (ja kui morf info võib olla ka tavakeele sõnaosa), tuleb päring hoolega läbi mõelda. Näiteks partitiivi märgend part võib olla ka tavakeele sõna. Kui tahta seda otsida partitiivi märgendina, võib otsida näiteks selle järgi, et see märgend on tavaliselt morf info lõpus, nt

○ part ○//

Morf analüüsist – mõned detailid

Üldiselt on järgitud eesti keele grammatikates esitatud sõnaliike. Võrreldes EKG-ga on siiski mõned erinevused, nt

- Adverbid – märgendatud on vaid üks adverbide klass _D_, pole jagatud eraldi afiksaal-, modaal- või proadverbideks nagu grammatikakirjelduste viimasel ajal kombeks.
- Adpositsioonid – kuna kaassõnad pole alati selgelt üheselt määratletavad (mingi osa on pidevalt grammatiseerumas, mistõttu palju kasutusjuhtumeid, kus on raske hinnata, kas pigem on tegu nimisõna või kaassõnaga), on otsus tehtud konstruktsiooni järgi: käändsõna genitiivis + teatud vorm või partitiivis + teatud vorm --> adpositsioon
- partitsiibid on märgendatud sõltuvalt nende süntaktilisest funktsioonist lauses. Kui partitsiip paikneb nimisõna ees (st on täiend), on ta üldjuhul adjektiiviks (_A_).
- suulise kõne tekstidele on lisatud sõnaliik *partikkel* märgendiga _B_

Märk = eraldab tuletusliited (peamiselt sõnaliiki muutvad tuletusliited) tüvest,
nt sikutamisega sikuta=mine+ga //_S_ com sg kom //

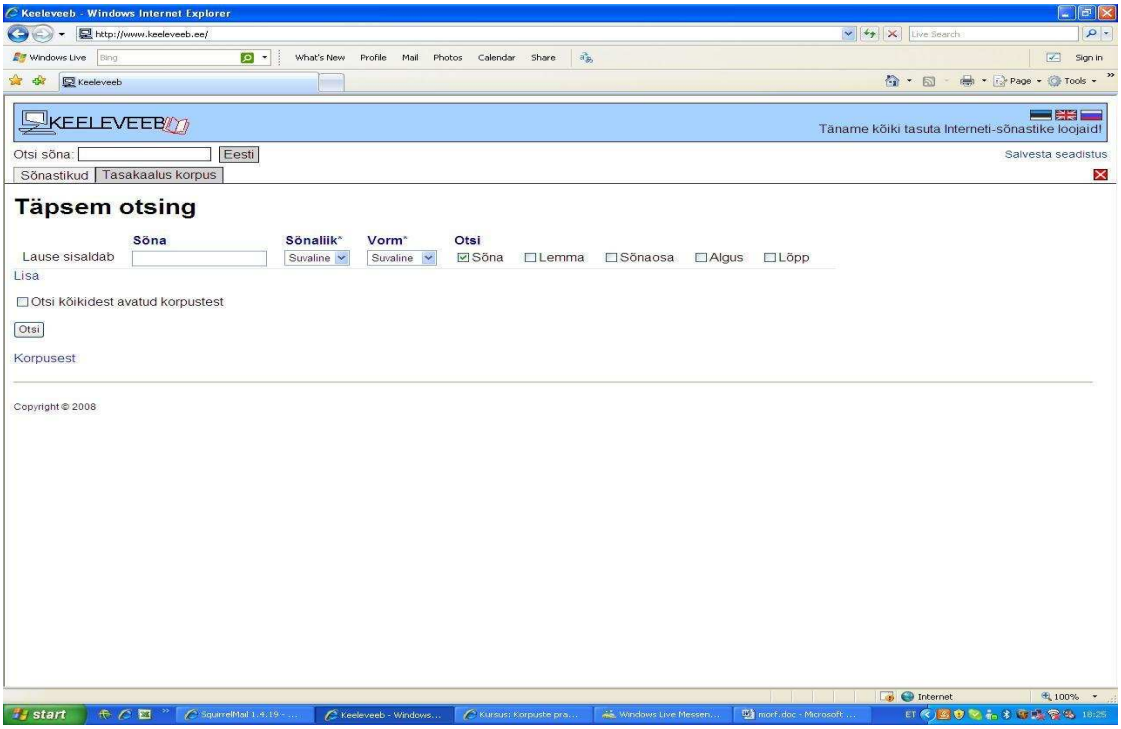
Täpsemalt vaata märgendamispõhimõtete kohta **H.-J. Kaalep, K. Muischnek, K. Müürisep, A. Rääbis, K. Habicht 2000**. *Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemusest.* Keel ja Kirjandus 9, lk 623-633, pdf: http://www.cl.ut.ee/yllitised/kk_2000.pdf

Keeleveebi päring morf. info põhjal

Keeleveeb on OÜ Filosofti lehekülg, kuhu on koondatud väga palju eesti keele ressursse, millest saab teha ka ühispäringuid. Kuna tegijad on osalt samad, on keeleveebis võimalik teha korpuspäringuid ka kirjakeele korpuse erinevatest allkorpustest. Uuem morf info põhjal toimiv otsingumootor paikneb samuti selle leheküljel. Siin on morfanalüsaatoriga automaatselt märgendatud tekstid, mis on ka automaatselt (st statistikapõhiselt) ühestatud. Võrreldes kirjakeele korpuse lehel olevate morf. märgendatud tekstidega võib selles otsimootoris rohkem ette tulla vigu, sest neid tekste ei ole lingvistid kontrollinud. Hinnanguliselt 3% analüüsidest on antud kontekstis valed, s.t. sõna algvorm, sõnaliik või grammatiline kategooria on määratud valesti. Analüüsiga seotud probleeme on kirjeldatud Heiki-Jaan Kaalepi ja Tarmo Vaino artiklis " Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis " kogumikus "Arvutuslingvistikalt inimesele", Tartu 2000, lk 87-99.

Samas on selle suureks plussiks see, et 1) morf märgendatud tekstide/sõnede hulk on oluliselt suurem kui kirjakeele korpuse lehel esitatutel, 2) on otsinugmootor oluliselt mugavam kasutada ning tekste saab lugeda ka ilma morf märgenditeta, st ka väljund on selgem ja mugavam edaspidi kasutada.

Morf märgendatud tekstidest otsimiseks ava www.keeleveeb.ee, sealt vali korpus (või mitu); sobivatel korpustel klõpsamistega avad need ning üles serva jäävad viited neile. Seejärel vali üks neist korpustest ning vajuta *täpsem otsing*. Nüüd leiad erinevad väljad, millele sisesta vajalik info.



NB!
- täitma ei pea kõiki välju
- korraga saad otsida ka kõigist avatud korpustest (tee linnuke " Otsi kõikidest avatud korpustest")
- päringu vastuseks v astuseks väljastatakse maksimaalselt 200 lauset. Kui päringule vastavaid lauseid on rohkem kui 200, siis väljastamiseks tehakse sobivate hulgast valik juhuslikult, kusjuures kahel järjestikusel päringul ei pruugi see valik olla samasugune.
- Märgenduse vaatamiseks tuleb soovitud sõnal klõpsata.
- Korraga saad otsida ka lauseid mitme tunnuse põhjal (vajuta lingile *Lisa*)

Väljade järgi:

sõna - siia kirjutatakse otsitav sõna algvormis (lemma), nt *olema, elama, sõber*

sõnaliik - sõnaliigid on samad, mis kirjakeele korpuse morfoloogiliselt märgendatud tekstides, vt http://www.cl.ut.ee/korpused/morfliides/seletus

vorm – verbivormid on esitatud üldiselt lõppude ja tunnuste järgi, käändsõnade puhul on käändelühendid

otsi – määrad, mis väljalt otsitakse. Näiteks kui väljale *sõna* on sisestatud *valitsus*, väljastatakse laused, kus tekstis ongi sõna *valitsus* algvormis. Kui tahta otsida sõna *valitsus* mingis teatud käändes, nt mitmuse nimetavas (pl n), tuleb linnuke teha hoopis kasti *lemma*, sest siis otsitakse sõna *valitsus* lemma väljalt (morf kirjelduses).

Kui otsida mingit sõnaosa, siis tuleb teha linnuke kasti *sõnaosa*. Nt järjendi *mata* leidmiseks tuleb väljale *sõna* kirjutada *mata* ning linnuke teha kasti *sõnaosa*.

Ülesanded ja vastused

NB! Päringu selgitustes kasutatud märk ○ tähistab tühikut!

1. Otsi tekstist eitussõna *ei*. Kuidas see on märgendatud?

Eitussõna *ei* võib kirjakeele korpusest otsida lihtsalt märgijada põhjal, arvestades, et enne ja pärast seda on tühikud.
○ei○
Nagu näha, märgendatakse eitussõna verbina. Kuna tänapäeva keele seisukohalt ta siiski verb ei ole, on ta muudes korpustes enamasti teistsuguse märgendi saanud (nt murdekorpuses on tal oma sõnaliik).

2. Otsi verbide *hakkama, saama, minema, võima* kasutusjuhtumeid.

Et otsida kõiki selle verbi vorme, tuleks otsingumootrisse sisestada sõna põhivorm nii, nagu seda korpuse märgenduses tehtud on, seega
○*hakka*

Võib täpsustada, et *hakka* oleks kindlasti tüvi, millele järgneb mõni muutevorm. Muutevorm lisatakse tüvele plussiga, seega võiks sisestada hakka+ Kuna pluss on ühtlasi erisümbol (märgib, et eelnev sümbol kordub üks või enam korda), otsitakse tegelikult välja kõik märgijadad *hakka, hakkaa* jne. Meil oleks aga vaja kõiki vorme, milles on plussiga märgitud, et järgneb lõpp. Selleks tuleks plussi erisümboli staatus kustutada, seda teeb nö vastukaigas \
○*hakka*\+

Kui *hakkama*-verbi puhul piisab ka lihtsalt tüve otsimisest (*hakka*), et ridamisi õiged vastuseid päringule saada, siis näiteks *saama* ja *võima* puhul on juba olulisem, et oleks märgitud, et tegu on tüvega, millele järgnevad lõpud:
○*saa*\+
○*või*\+

Verbi *minema* puhul tulevad morfoloogilselt märgendatud tekstide kasutamise eelised eriti selgelt välja, on tal ju kaks tüve (*mine-, lähe-*), nüüd piisab aga *ma*-infinitiivi tüve kasutamisest:
○*mine*\+

Verbi *võima* puhul on kasulik märkida juurde, et sõna peaks olema kindlasti verb (vastasul juhul leitakse ka sidesõna *või*, nimisõna *või*). Verbi märgend on _V_. Tuleb arvestada, et tüve *või+* ja märgendi _V_ vahel võib olla muudki, seepärast võiks siin kasutada välistamistaktikat ja päringut täpsustada nii, et *või+* ja _V_ vahel võib olla

misiganes, v.a alakriips _, millele omakorda järgneb _V_. Sellega välistame, et mõne teise sõnaliigi tunnus sinna vahele jääb:
○või\+[^_]*_V_

3. Kuidas on märgitud *gi/ki*-liide?
Otsi *gi/ki* + tühik:
[gk]i○

4. Otsi *nud*-partitsiibi kasutusjuhtumeid. Mis on kõige tavalisem *nud*-partitsiibi kasutusala korpuste põhjal?

Et leida *nud*-partitsiibivorme ja mitte muidu *nud*-järjendeid, tuleks otsingut piirata nii, et *-nud* võib esineda vaid peale tüve:
\+nud

Päringu tulemusi vaadatates saab selgeks, et nud-partitsiibi puhul on eraldi märgendi saanud selle eri funktsioonid.

Täiendina esinev *-nud* on märgitud nagu omadussõna (sama kehtib ka muude partitsiipide kohta):
Mõned mõni+d //_P_ pl nom // lagunenud **lagunenud+0 //_A** pos_ // sängid säng+d //_S_ com pl nom //

Eituskonstruktsiooni osana on *nud*-partitsiip ta saanud oma märgendi:
Ta tema+0 //_P_ sg nom // ei ei+0 //_V_ aux neg // tahtnud taht+nud **//_V_ main indic impf ps neg //**

Liitaegades:
Olin ole+in //_V_ aux indic impf ps1 sg ps af // uurinud uuri+nud **//_V_ main partic past ps //** niisugust nii_sugune+t //_P_ sg part // pilti pilt+0 //_S_ com sg part // , , //_Z_ Com //

5. Otsi *v*-partitsiibi kasutusjuhtumeid (nt *lugev*, *seisev*).

Loogiline oleks kasutada *v*-partitsiibi märgendit _V_ main partic pres ps
Nii märgendatud partitsiibivorme on aga mõni üksik, enamus on kuidagi teisiti märgendatud.

Üks võimalus olukorda lahendada otsida *v*-lõpulisi sõnu (päring v○) ja vaadata täpsemalt, kuidas on v-partitsiip märgendatud (üldiselt on sõnaliigiks _A_).

6. Otsi partikleid suulise kõne tekstidest.

Partikli märgend on _B_

7. Otsi käskiva kõneviisi eitusvorme (*ära tee, ärge tehke, ärgem tehkem~ärme teeme~ärme tee, ärgu tehku*).

Kuna nii käskib kõneviis kui eitus on morf infos olemas, tuleks ülesande lahendamiseks kõigepealt silmitseda käskiva kõneviisi eitavate vormide märgendeid.

_V_ main imper pres ps2 sg ps neg	Põhiverb imperatiiv preesens 2. pööre singular personaal negatiiv	Verb main imper present second singular active negative	Vmmp2s-ay	68	(ära) loe
_V_ main imper pres ps3 sg ps neg	Põhiverb imperatiiv preesens 3. pööre singular personaal negatiiv	Verb main imper present third singular active negative	Vmmp3s-ay	2	(ärgu) lugegu
_V_ main imper pres ps1 pl ps neg	Põhiverb imperatiiv preesens 1. pööre pluural personaal negatiiv	Verb main imper present first plural active negative	Vmmp1p-ay	0	(ärgem) lugegem
_V_ main imper pres ps2 pl ps neg	Põhiverb imperatiiv preesens 2. pööre pluural personaal negatiiv	Verb main imper present second plural active negative	Vmmp2p-ay	23	(ärge) lugege
_V_ main imper pres ps3 pl ps neg	Põhiverb imperatiiv preesens 3. pööre pluural personaal negatiiv	Verb main imper present third plural active negative	Vmmp3p-ay	3	(ärgu) lugegu
_V_ main imper pres imps neg	Põhiverb imperatiiv preesens impersonaal negatiiv	Verb main imper present passive negative	Vmmp---py	1	(ärgu) loetagu

Tabelist näeme, et kõigil juhtudel on sarnane märgendi algus _V_ main imper pres ja lõpp neg, nende vahel on muud märgendid. Seega tuleks otsida märgijada, kus algus ja lõpp on teada, vahepealne osa aga varieerub. Vahepealne osa peab aga kindlasti olema sama vormi kirjeldus, mitte mõne muu vormi kirjeldus, seega võiks piirata nii, et see ei ületaks märgendi lõpu piiri, milleks on //
Seega võiks päring välja näha selline:

_V_ main imper pres[^/]* neg

Sama päring keeleveebist – tuleb otsida ükshaaval igat vormi:
sõnaliik V
vorm neg ge (järgmisel otsingu neg gem jne)
linnukest kastidesse *sõna, sõnaosa, algus, lõpp* pole vaja

8. Otsi resultatiivsust väljendavat konstruktsiooni *saama* (*saan, saad, said, sai* jne) + *tud*-partitsiip (*sai käidud, sain tehtud*).

Selles ülesandes tuleb arvestada sõnatüvega saa\+ (vt ül 2) ja *tud*-partitsiibiga (+tud), aga tuleb arvestada, et nende vahel võib olla muidki sõnu. See, mida nende vahel ei tohi olla, on kirjavahemärgid, sest kirjavahemärgid jagavad nad eri osalausetesse. Kirjavahemärgi märgend on _Z_, seega võiks päring välja näha nii:
○saa\+[^Z]*_V_○main○partic○past○imps

See otsing teeb selle vea, et hõlmab ka juhtumeid, kus *saama*-verb ise on *tud*-partitsiibi vormis:
Vahel vahel+0 //_D_ // , , //_Z_ Com // kui kui+0 //_J_ sub // olengi ole+ngi //_V_ aux indic pres ps1 sg ps af // iseenda ise_enese+0 //_P_ sg gen // ja ja+0 //_J_ crd // oma oma+0 //_P_ sg gen // vanemate vanem+te //_S_ com pl gen // saatusest saatus+st //_S_ com sg el // kõnelnud kõnele+nud //_V_ main partic past ps // , , //_Z_ Com // on ole+0 //_V_ aux indic pres ps3 sg ps af // minust mina+st //_P_ sg el // ikka ikka+0 //_D_ // valesti valesti+0 //_D_ // aru aru+0 //_S_ com sg part // saadud saa+dud //_V_ main partic past imps // - — //_Z_ Dsh //

Seega tuleks otsingut modifitseerida nii, et ta otsiks *saama*-verbile järgnevat partitsiipi kaugemalt kui sama sõna kirjeldusest. Selleks lisame otsingusse piirangu, et ta otsiks partitsiipi järgmisest märgendatud sõnast alates, st siis järgmisest plussist alates (pluss on tüve ja lõpu vahel).
○saa\+[^Z]*\+[^Z]*_V_○main○partic○past○imps

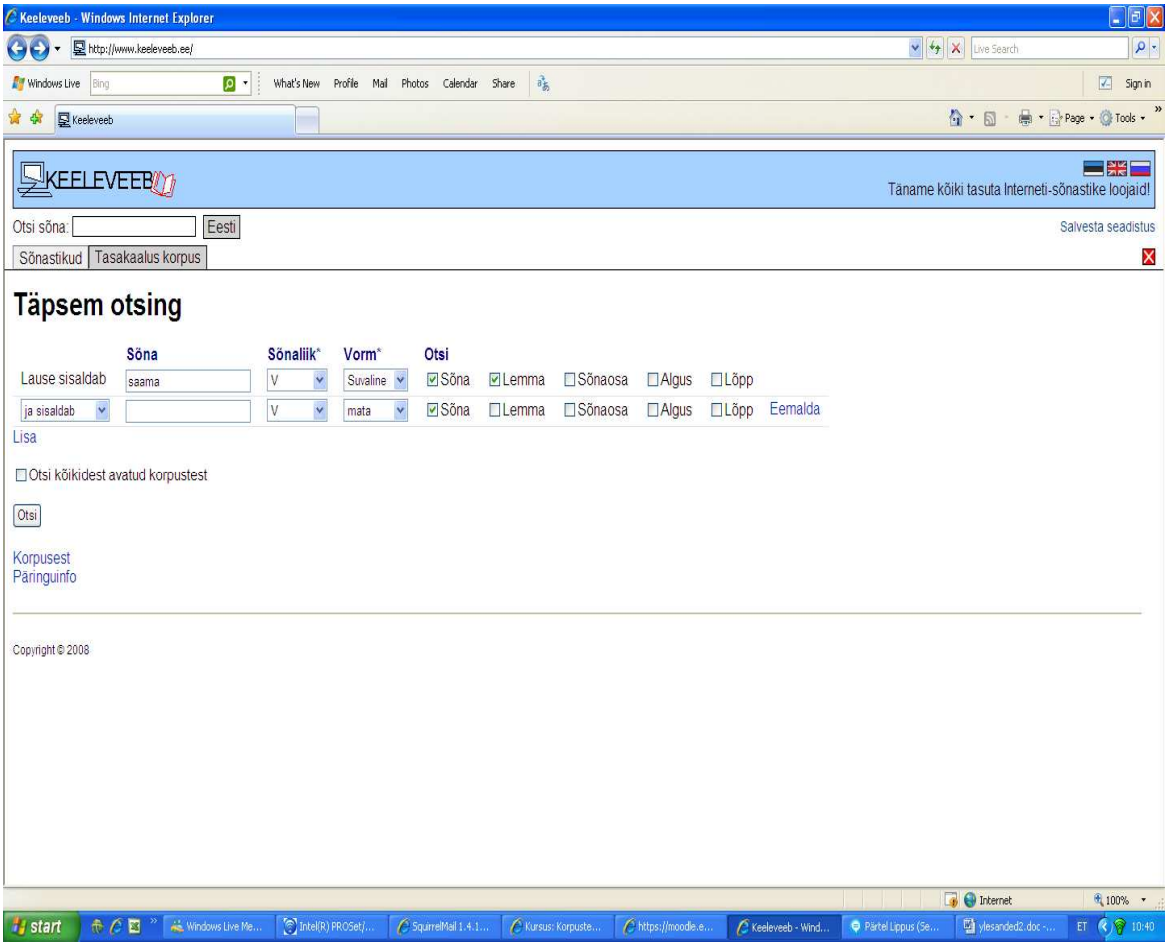
Et hõlmata kõiki korpuses leiduda võivaid lauseid, tuleks arvestada ka teistsuguse sõnajärjega, nii et partitsiibile järgneks enne kirjavahemärke *saama*-verb.
_V_○main○partic○past○imps○[^Z]*○saa\+

Need kaks päringut võib ka ühendada, selleks oli nn toru | , mis tähendab, et otsitakse kas torule eelnevat või järgnevat märgijada:

○saa\+[^Z]*\+[^Z]*_V_○main○partic○past○imps|_V_○main○partic○past○imps○[^Z]*○saa\+

Sama päring keeleveebist:

esimesse ritta kirjutada *saama* ja tähistada, et tegu lemmaga
lisada teine rida (valik: *ja sisaldab*), sinna kirjutada vormi kirjelduseks *mata*



Nagu näeme, on selle otsingu puuduseks asjaolu, et konstruktsiooni otsitakse kogu lausest, mitte ühe osalause piirest, ning täpsemalt piirtleda otsingut ei saa.

9. Otsi progressiivkonstruktsiooni *olema* +*-mas* (*on valmimas*).

See ülesanne tuleks lahendada eelnevaga analoogiliselt: otsida *olema*-tüve (siin pole selge, kas see on märgendatud põhiverbiks või abiverbiks, kindlam on seega tüve kasutada):
○ole\+
Sellele peaks järgnema *mas*-vorm, aga kindlasti sama osalause sees, seega kasutame eitust: olema verbile võib järgneda misiganes ja ükskõik kui palju, v.a kirjavahemärk (mille tähis on _Z_), seega:
○ole\+[^Z]*

mas-vormi morf lühend on: _V_ main sup ps in

Need kokku pannes:
○ole\+[^Z]*_V_○main○sup○ps○in

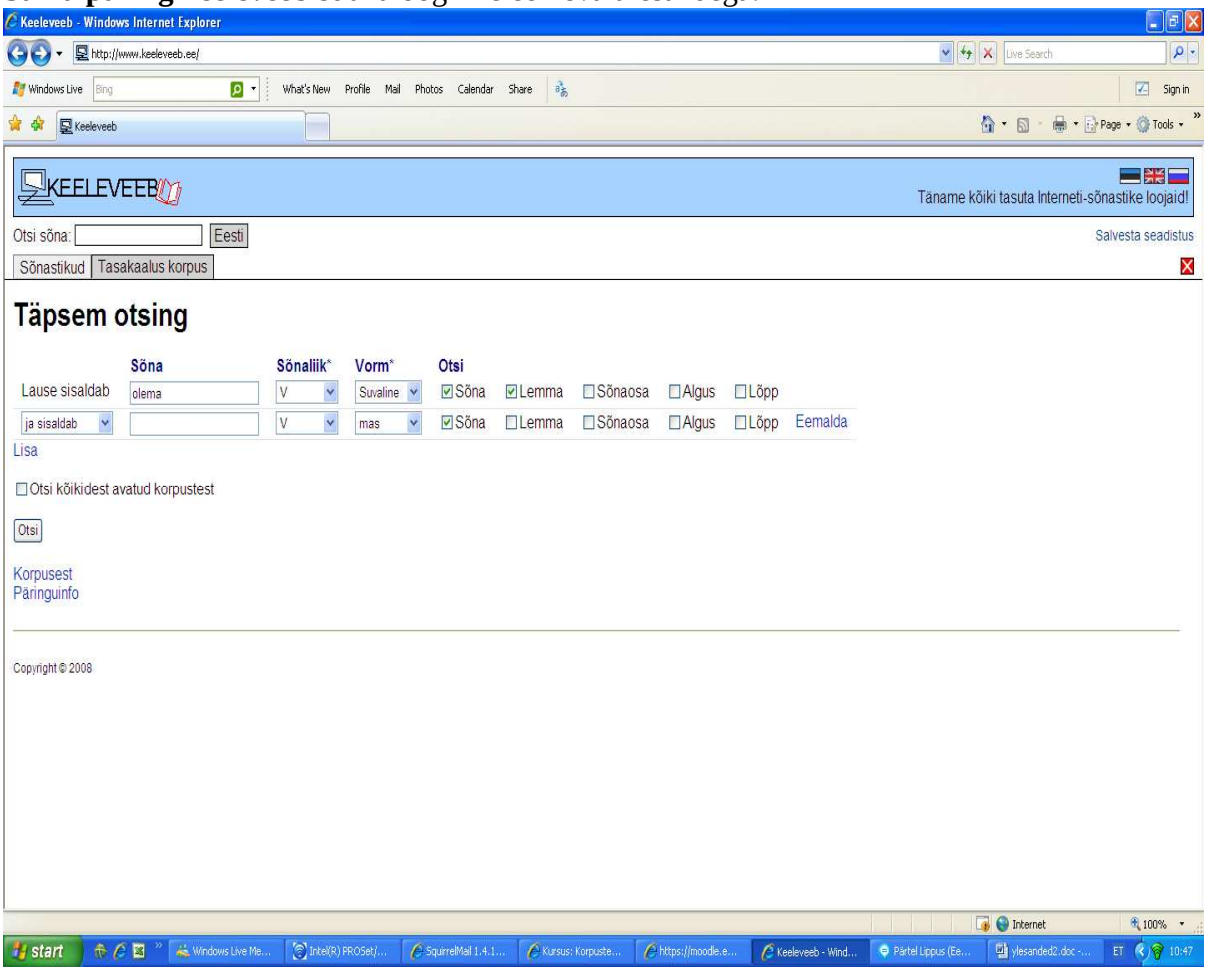
Ka see päring teeb kaks viga: 1) otsib ta ka ainult *olemas*-vorme, 2) ei arvesta teistsuguse sõnajärjega.

Kasutame samu võtteid, mis eelmises näites:
1) lisame nõude, et *mas*-vorm peab olema vähemalt järgmises sõnas, mitte *olema*-sõnas:
○ole\+[^Z]*\+[^Z]*_V_○main○sup○ps○in
2) Teeme teise päringu vastupidise sõnajärjega, nii et *mas*-vorm oleks enne *olema*-verbi:
_V_○main○sup○ps○in[^Z]*○ole\+

Muidugi võib need kaks otsingurida nüüd ka kokku panna ja otsida kõik laused korraga välja:
○ole\+[^Z]*\+[^Z]*_V_○main○sup○ps○in|_V_ main○sup○ps○in[^Z]*○ole\+

Tulemus on nüüd selline, et uurimistööks vajaliku leiab siit kergesti üles, ei ole liiga palju "prahti" hulgas.

Sama päring keeleveebist analoogiline eelneva ülesandega:



4. Vana kirjakeele korpus http://www.murre.ut.ee/vakkur

Vana kirjakeele korpus hõlmab vanemaid kirjalikke eestikeelseid tekste alates 13. sajandist. Järgnevast kursuse materjalist leiate vana kirjakeele korpuse tutvustuse, mis on koostatud korpuse töö juhi Külli Habichti materjalide põhjal. Sellele tuleks lisaks lugeda 2004. a Keeles ja Kirjanduses ilmunud ülevaateartiklit V.-L. Kingisepp, K. Prillop, K. Habicht 2004. "Eesti vana kirjakeele korpus: mis tehtud, mis teoksil" - Keel ja Kirjandus, 4, 272-280.

Üks eesti kirjakeele ajaloo tähtsamaid väljaandeid - 1739. a Piibel - on saadaval aga Eesti Keele Instituudi leheküljel, vt
http://portaal.eki.ee/piibel/index.php?tekst=tutv_pbel

Seal on ka lõunaeestikeelne 1686.a Vastne Testament
http://portaal.eki.ee/piibel/index.php?tekst=tutv_wast

Korpuse tutvustus

Vana kirjakeele korpust (VAKKUR) on Tartu ülikoolis loodud alates 1995. aastast. Korpus on mõeldud eelkõige keeleuurijatele ning annab võimaluse keele diakrooniliseks uurimiseks.

Spetsiifika

Võrreldes eesti kirjakeele korpusega on vana kirjakeele korpuse loomine oluliselt enam aega ja teadmisi nõudev tegevus. Eri ajastute ja eri autorite keel on küllalt erinev (st suur varieeruvus), samuti on vaja teadmisi nii gooti kirjast kui saksa keelest, seetõttu ei saa seda tööd teha päris igaüks. Näiteid selle kohta, millised on need käsikirjad ja materjalid, millega tuleb töötada, leiate korpuse kodulehelt:
http://www.murre.ut.ee/vakkur/Gooti/pildid.htm. Käsitsitöö ja keerukuse tõttu ei ole vana kirjakeele korpus ka mõõdetav miljonites nagu tänapäeva kirjakeele korpus, vaid seda on kokku umbes 2,2 miljonit tekstisõna (2009. a alguse seisuga).

Korpuse ülesehitus

Korpus koosneb kolmest suuremast allosast:

• 1995. a alustatud vanimate eestikeelsete tekstide lauskorpus (16. sajandist ja 17. sajandi esimesest kümnendist ka käsikirjad) kuni 1660. aastateni. Lauskorpus tähendab seda, et korpusesse on hõlmatud kõik sellest ajavahemikust säilinud eestikeelsed tekstid. Kokku on selle korpuseosa maht 900 000 tekstisõna.

• 2002. aastast 18. sajandi tekstide valikkorpus. Umbes 800 000 tekstisõna.

Korpuse nende kahe allosa sisu saab vaadata ja mõningaid tekste lugeda siit: http://www.murre.ut.ee/vakkur/Korpused/korpused.htm

• 2005. aastast on loodud 19. sajandi esimese poole tekstide valikkorpust, milles on umbes 500 000 tekstisõna. Selle korpuseosa kohta saab lugeda ja teha päringuid siin: http://www.murre.ut.ee/vakkur/Korpused/Kwic2/paring19.htm

Mida korpusest otsida saab?

Praegu päringud vaid märgendamata tekstist (vt lingid eespool). Päringusüsteem märgendatud tekstist on alles arendamisel ja katsetamisel (Külli Prillop).

Vanemad tekstid (kuni 18. saj)

Otsingumootor: http://www.murre.ut.ee/vakkur/Korpused/Kwic2/paring.htm

NB! Otsimootori kasutamisel tuleb silmas pidada, et otsimootorisse sisestakse terve sõna, mitte sõnaosa. Kui on vaja otsida ainult sõnaosa järgi (nt käändelõpu järgi), märkige erisümboleid kasutades ka sõna algusosa päringuväljale, nt .*le . Kui on vaja otsida sõna alguse järgi, märkige algus ning lõpus kasutage erisümboleid: po[oh]l.*

Erisümbolid on samad, mis kirjakelele korpuses, ent siin on nende hulk piiratum, vt selgitusi päringute juures.

Probleemid: kasutaja peab arvestama vanimate tekstide puhul ebakorrapärase kirjaviisiga (võõrtähed, tilde (~) nasaalide asemel jm) ning alates 17. sajandi lõpust vana kirjaviisi eripäradega (nt pikkade ja lühikeste häälikute märkimine tänapäevasest erinev). Kõiki variante pole võimalik ennustada, nt *poohomene* ‘poomine’.

Nt sõna *pool* variandid G. Mülleri jutlustes (1600— 1606): *pohl*, *poel* – siin piisaks päringu *po[he]l*, ent vrd *tuul* – päring *tu[he]l* ei anna variante *twl*, *thul*, *tul*, st enamik variante jääb leidmata. Alles päring *th*[uw]h*l* annaks ammendava vastuse.

Kasutaja peab orienteeruma ka vana kirjakeele vormimoodustustavades. Näiteks G. Mülleril varieeruvad vormid *anda* , *andada* ; *istwat* , *istuwat* ; *hüppas* , *hüppis* . J. Rossihniusel (1632) varieeruvad vormid *minnenut* , *minnut* , *lahenut* , *lennut* ‘läinud’. Varieerumine on pigem reegel kui erand.

Täpselt otsida saab vaid seda infot, mida teatakse või osatakse ette näha, seepärast on väga oluline leida infot, kuidas üks või teine sõna varasemates tekstides üldse välja võib näha. Praegusel hetkel aitavad selles kirjus pildis orienteeruda trükitud abivahendid, nt konteksti ja lisanäiteid saab otsida trükitud sõnastikest (vanimate tekstide sõnastik 1997, Müller 2000, Stahl 2002, Rossihnius 2002, vt täpsemaid viiteid siit: http://www.murre.ut.ee/vakkur/Yllitised/yllitised.htm). Samuti aitavad veebis olevad sõnastikud, vt http://www.murre.ut.ee/vakkur/Korpused/veeblug.htm

Vanemate tekstide, Mülleri jutluste, Turu käsikirja ning Rossihniuse kirikukäsiraamatute kohta olemas ka **märksõnastatud tekstid**, mis on ka eespool mainitud veebisõnastike alusmaterjal, vt http://www.murre.ut.ee/vakkur/Korpused/Kwic/paring.html.

Märksõnastatud tekstid aitavad ületada kirjaviisi varieerumise (ning sealt saab kirjaviisi ka muidugi kontrollida). Märksõnastatud tekstid sisaldavad märksõna (tänapäevasel kujul, käändsõna nimetavas käändes, pöördsõna *ma*-infinitiivi vormis) ning sõnaliigi infot. **Korpuse tekstid ei ole sõnaliikide ja tähenduste osas täielikult ühestatud**, st kui mingi sõna on kasutatav mitmes sõnaliigis või tähenduses, ei ole iga konkreetse kasutusjuhu puhul otsustatud, millises ta just parasjagu on. (Nt *pärast* taga on nii märgendid ADP kui ka ADV).

19. sajandi tekstid

http://www.murre.ut.ee/vakkur/Korpused/Kwic2/paring19.htm

19. sajandi tekstide puhul on kirjaviis ühtlasem (vana kirjaviis) ning varieerumine on väiksem.

19 .saj tekstidest otsides tuleb arvestada sellega, et oletuslikult otsitakse tervet sõna (nagu varasemate tekstide puhulgi), ent siin on võimalik ka otsida vaid sõnaosa järgi, kui võtta linnuke ära kastist *Otsi tervet sõna*. Siis on võimalik otsida ka ainult nt käändelõpu järgi, ilma algusosa sisestamata.

Vana kirjakeele korpuse morfoloogiline märgendamine

Ka vana kirjakeele tekste on hakatud morfoloogiliselt märgendama. Selleks on loodud abivahend VAKKER (autor Külli Prillop). Tekstid on XML-formaadis.

Märgendatakse jooksvat teksti, sest loendi märgendamisel võib tekkida vigu vormihomonüümia tõttu: *kena – kena , kääna olema – olema* (v), *olemine* (s) *liiwa – liiva , leiva* Mida sagedam sõne, seda suurem veaoht. Nt Mülleril *ollema* 171 korda, neist ühel korral substantiivsena tähenduses ‘olemine’.

• Programm soovitab märksõna, sõnaliiki ja grammatilist infot. Selleks on kasutusel:

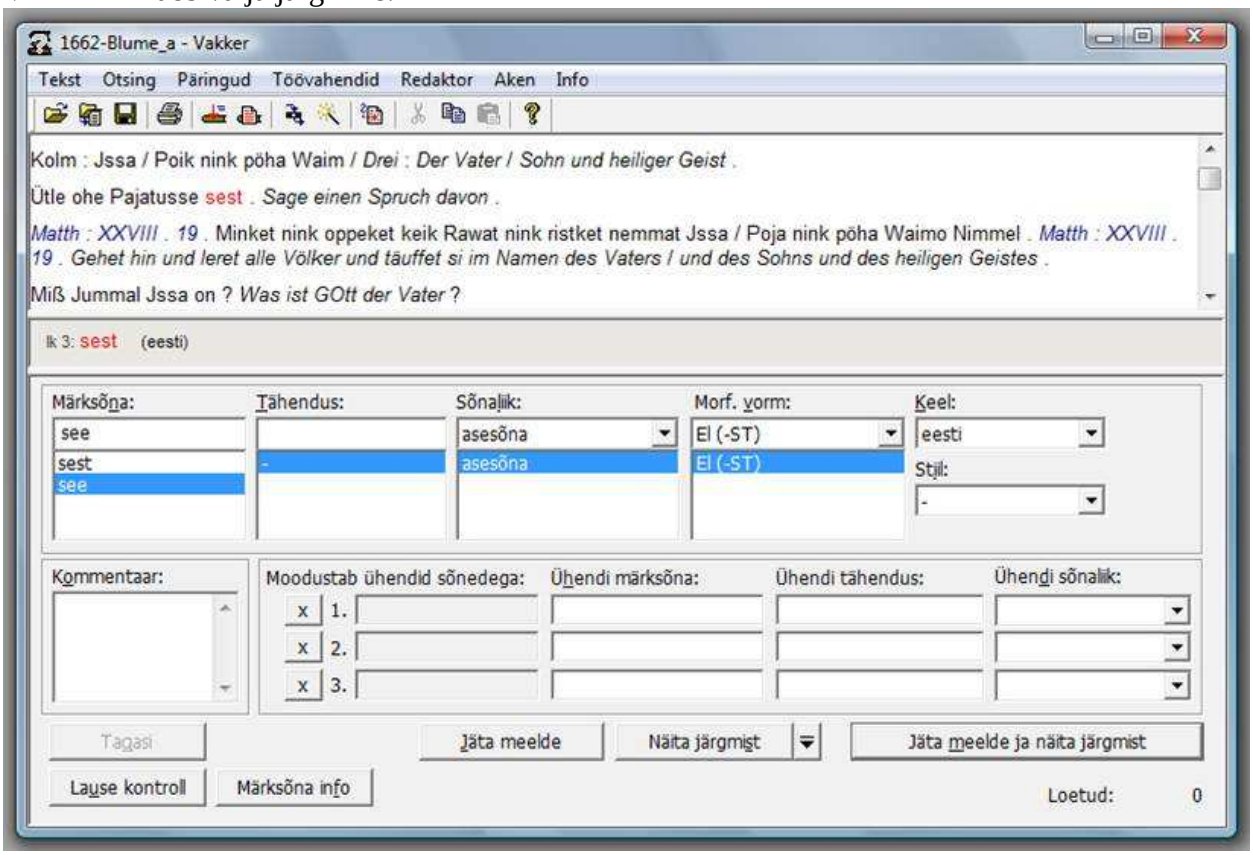
1) **ESTMORF** (morf analüsaator) + teisendusreeglid vana sõna umbkaudseks tänapäevastamiseks, nt *e > ee , e > ä* .

2) Juba lisatud info talletatakse **abisõnastikku** (kui sõna tuleb tekstis ette mitmendat korda, saab vajaliku info abisõnastikust).

3) Umbkaudne otsing **sõnastikust** (ei arvesta sõnalõppe), nt kui on olnud sõna *oppema* , siis pakub õige lemma ka vormile *oppenut* .

4) **Grammatilised “lisateadmised”,** nt kui tegemist vokaallõpulise nimisõnaga (tänapäeval aga lõpus konsonant) ja selget käändetunnust pole, siis pakub genitiivi vormi.

VAKKER näeb välja järgmine:



5. EKI korpused, õppijakeele korpused.

Selle teema alla on koondatud infot mõningate väiksemate ja/või spetsiifilisemate korpuste kohta. Loe iga korpuse kohta käivaid materjale.

EKI tekstikorpus http://portaal.eki.ee/corpus

Emotsionaalse kõne korpus (EKI) http://193.40.113.40:5000/ , loe lisaks: Altrov, Rene 2008. Eesti emotsionaalse kõne korpus: teoreetilised toetuspunktid. Keel ja Kirjandus, 4, 261 - 271.

Eesti vahekeele korpus (Tallinna ülikool) - korpuse lühitutvustus + kasutajaliides http://evkk.tlu.ee/wwwdata/what_is_evk . Vahekeele korpust tutvustav artikkel: P. Eslon, H. Metslang "Õppijakeel ja eesti vahekeele korpus" , Eesti Rakenduslingvistika Ühingu aastaraamat3, 2007, lk 99-116.

Eesti lastekeele korpus (Tallinna ülikool): R. Argus "Eesti lastekeelekorpuse morfoloogilisest märgendamisest". - Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine / Toim. P. Eslon. Tallinna Ülikooli eesti filoloogia osakonna toimetised 9. Tallinn: Tallinna Ülikooli Kirjastus, 2007, lk 65-86.

6. Spontaanse kõne foneetiline korpus

Korpus on loodud foneetiliste uurimuste jaoks.

Praegusel hetkel on korpuse üks kasutusvõimalusi kasutada veebipõhist otsimootorit, millest otsimine on siisk iseotud teatavate piirangutega (kontekst 2 sek, ei saa otsida kõigilt märgendustasanditelt). Otsimootor paikneb http://www.murre.ut.ee/otsing/ekskfk.php.

Eesti keele spontaanse kõne foneetiline korpus

(Koostatud Pärtel Lippuse ja Pire Terase materjalide põhjal)

Eesti keele spontaanse kõne foneetilist korpust luuakse Tartu ülikoolis alates 2006. aastast. Korpus pakub materjali eelkõige foneetika uurijatele. Eesmärk on luua spontaanse kõne foneetiliselt märgendatud korpus, mida saab kasutada eesti keele häälduse põhiparameetrite analüüsimisel ning eesti keele kõnesünteesi ja kõnetuvastuse ülesannete täitmisel. Selleks tehakse spontaanse kõne kõrge kvaliteediga salvestusi ning salvestatud kõne märgendatakse foneetiliselt erinevatel märgenduskihtidel (sõna, silp, häälik jne).

1. Korpuse ülesehitus

Korpuse põhiosa moodustavad kaasaegse eesti keele helisalvestised. Salvestised on transkribeeritud ja segmenteeritud programmiga Praat. Selle käigus lisandub helisalvestisele tekstifail, mis sisaldab kogu transkribeeritud-märgendatud info (TextGrid), tänu millele on juba võimalik korpusest otsida kui tekstifailist, st litereeringu ja märgenduse põhjal.

Lisaks salvestistele ja TextGrididele on korpuses ka info kõnelejate ja salvestiste kohta.

2. Korpuse maht ja kõnelejate valiku kriteeriumid

Korpuse koostamise esimeses etapis on kavas lindistada 40 kõnelejat. Kavandatud on umbes pool tundi kõnet igalt keelejuhilt, seega korpuse kogumahuks on planeeritud 20 tundi. Tegelikkuses on salvestuste kestus varieerub vahemikus 20-50 minutit ja mitu keelejuhti osalevad mitmes salvestuses.

Kõnelejad on eri vanuses (ligikaudu 12 kõnelejat teismelised ja 20ndates, 8 kõnelejat 30ndates, kaheksa kõnelejat 40ndates, 12 kõnelejat 50ndates ja vanemad) ja eri soost (pooled mehed, pooled naised). Proovitakse leida kõnelejaid, kel oleks erinev piirkondlik ja sotsiaalne taust. Ülevaate korpuse hetkeseisust leiab siit . Iga keelejuht täidab enda kohta taustainfot sisaldava ankeedi, milles ta annab ka nõusoleku, et tema kõne lindistusi korpuses kasutatakse. Keelejuhid kodeeritakse (nt 001_N, 002_M – keelejuhi

number_sugu). Kui üks keelejuht osaleb mitmel lindistusel, kasutatakse tema kohta sama koodi. Isikuandmeid kõrvalistele isikutele ei avaldata.

Salvestised

Korpuse tarvis lindistatakse spontaanseid argidialooge, kus keelejuhid vestlevad vabalt valitud teemadel. Lindistatakse (pool)spontaanseid institutsionaalseid monolooge ettekannete, loengute vms näol (nende puhul on siis tegemist ettevalmistatud, kuid mitte ette loetud tekstiga). Võimalusel tehakse salvestused Tartu Ülikooli ajakirjandusosakonna helistuudios, kus mõlemal kõnelejal on oma mikrofon ja kanal ega ole „segajaid“ (v.a esialgu võõras situatsioon). Lindistatakse otse arvutisse. Kui stuudiolindistus pole võimalik, siis lindistatakse keelejuhti nt tema kodus, kus on tingimuseks vaikne, liigse müraustata ruum. Ka siis on igal kõnelejal oma mikrofon ning lindistatakse otse arvutisse. Poolspontaanse institutsionaalse monoloogi puhul kasutatakse pea külge kinnitatavat mikrofoni ja digitaalsalvestajat.

Helisalvestised salvestatakse wav-formaadis lineaarsetena resolutsiooniga 16 bitti ja 44.1 kHz, signaale ei töödelda. Salvestatud failid saavad nimetuse vastavalt lindistusele ja kõnelejale (nt SKK002-001_N.wav – spontaanse kõne korpuse 2. lindistus-1. keelejuht_naine). Iga salvestise juurde kuulub tekstifailina salvestuse taustainfo (salvestamise aeg, tehnilised andmed jms).

Segmenteerimine ja märgendamine

Kõik helisalvestised segmenteeritakse ja märgendatakse (st transkribeeritakse, määratakse erinevate kõneüksuste piirid, lisatakse mitmeid märgenduskihte). Segmentimisel ja märgendamisel kasutatakse kõneanalüüsiprogrammi Praat (Paul Boersma ja David Weeninki poolt Amsterdami Ülikoolis väljatöötatud programm).

Segmenteerimise käigus leitakse erinevate üksuste (sõnad, häälikud, silbid) piirid ning lisatakse info, mis igal tasandil selles lõigus on. Sõna kirjutatakse tavalises ortograafias, häälikutasandil kasutatakse SAMPA transkirptsiooni.

Märgenduskihid on järgmised:

• **sõnad** (ortograafiline kirjaviis; siin ka üneemid, nt .ee);

• **häälikud** (SAMPA transkriptsioonis);

• **häälikustruktuurid** (CV) – teisendatakse häälikutest;

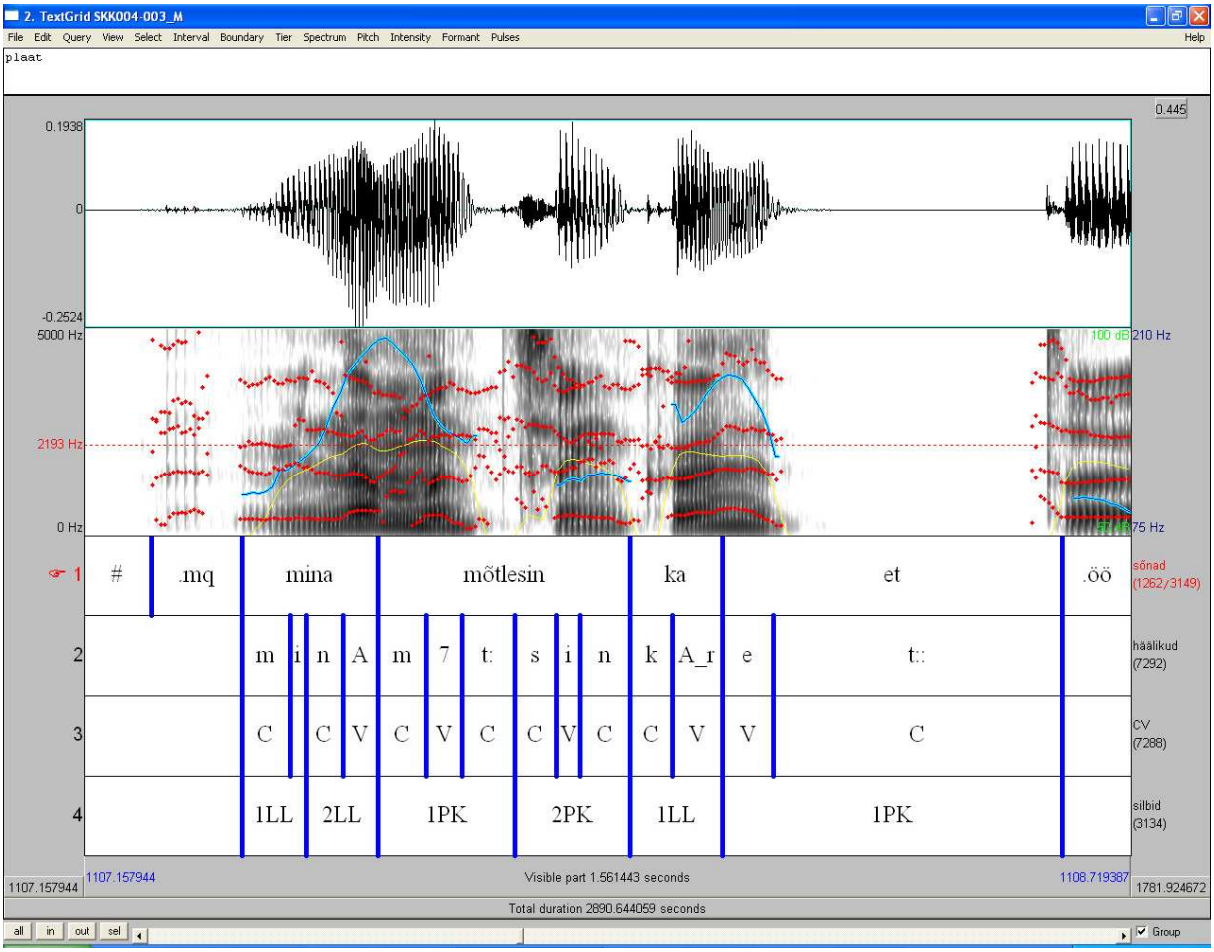
• **silbid** – LL (lühike, lahtine), PL (pikk, lahtine), PK (pikk, kinnine) + silbi järjekorranumber. Nt kau|ba|ma|ja – 1PL|2LL|1LL|2LL

• **taktid** – siin märgime pearõhku (1) ja kaasrõhku (2) ning väldet. Nt kava|lamale – 11|21, kavala|male – 11|21;

• **lausungid** (JUTT, PAUS, täidetud paus, hingamine).

Erinevate üksuste piirid ning eri tasandite info kirjutatakse TextGridi. TextGrid n-ö hoiab lisainfot helifaili küljes kinni: TextGrid sisaldab kõigi üksuste algus-jalõpuaega helifailis.

Järgnevalt on näide helifailist ja sellele vastavast märgendatud tekstilõigust nii, nagu see Praatis paistab. Detailsemat infot märgenduskihtide ning SAMPA transkriptsiooni kohta vaata http://www.murre.ut.ee/triip/margendus/.



Pildil on Praati aken, selle kõige ülesmises avas on helilaine, teises avas spektrogramm. Spektrogrammilt võib lugeda infot heli kõrguse, formantide, intensiivsuse jms. kohta Punased täpikesed näitavad formante, kollane joon intensiivsust, sinine joon põhitooni.

Järgnevad kihid on juba märgenduskihid. 1. kihis on sõna ortograafilises kirjaviisis, ent mõningase lisainfoga: punkt sõna ees tähendab, et tegu on kommentaariga, nt .sisse tähendab sissehingamist, .mq on mingi üneem või häälitsus, # tähendab pausi.
2. märgenduskiht märgib häälikuid, 3. kiht häälikustruktuure (konsonant või vokaal), 4. kiht iseloomustab silpe. Selles näites ei ole 5. ja 6. märgenduskihti.

Järgnevalt on sama lõik **TextGridis** (sõna tasandil):

intervals [1253]: lõigu ID,

xmin= alguspunkt helifailis (kaugus salvestuse algusest sekundites),

xmax= lõigu lõpp helifailis,

text= sõna tasandi märgendus või transkriptsioon.

```
  intervals [1253]:
          xmin = 1106.0904130775384
          xmax = 1106.3906876528156
          text = ".sisse"
      intervals [1254]:
          xmin = 1106.3906876528156
          xmax = 1107.2574038492296
          text = "#"
      intervals [1255]:
          xmin = 1107.2574038492296
          xmax = 1107.3923306587146
          text = ".mq"
      intervals [1256]:
          xmin = 1107.3923306587146
          xmax = 1107.5964390150343
          text = "mina"
      intervals [1257]:
          xmin = 1107.5964390150343
          xmax = 1107.9716545364654
          text = "mõtlesin"
      intervals [1258]:
          xmin = 1107.9716545364654
          xmax = 1108.1091457407683
          text = "ka"
      intervals [1259]:
          xmin = 1108.1091457407683
          xmax = 1108.6173245090158
          text = "et"
      intervals [1260]:
          xmin = 1108.6173245090158
          xmax = 1109.1418527356327
          text = ".öö"
```

Samamoodi on kirjeldatud ka muude märgendustasandite algus- ja lõpuajad.

Foneetikakorpusest otsimine

Foneetikakorpuse otsimootor paikneb aadressil
http://www.murre.ut.ee/otsing/ekskfk.php.

Sellega saab teha esialgseid päringuid korpuse materjali kohta. Veebipõhine otsimootor võimaldab otsida korpusest ühe sõna piires, vastuseks antakse 2-sekundiline helilõik ja selle märgendus. Lisaks võib endale alla laadida ka sama lõigu TextGridi, mida saab kasutada programmiga Praat.

Korpuse kõnelejate kaitseks on veebipõhisest otsimootorist välja jäetud isikunimed, seda nii helifailis kui märgenduses.

Otsimootorist saab praegu otsida vaid sõna tasandilt ortograafilisi sõnesid (st tekstisõnu, nagu nad tekstis on). Tähele tuleb panna, et kirjeldada tuleb terve sõna. Kui sõna algus või lõpp pole teada (st lõpus võib olla muutlõpp vms), võib kasutada erisümboleid. Erisümbolid on põhimõttelised samad, mis kirjakeele korpuses. Näiteks verbi *tulema* vormide otsimiseks võiks kasutada *tul.** Otsingumootori ja erisümbolite kohta vt lisaks http://www.murre.ut.ee/triip/otsingu-kasutamisjuhised/

Kui otsimootori võimalused osutuvad liialt piiratuks, on võimalik kasutada ka Praati otsimootrit. Selleks tuleb kirjutada korpuse administraatorile Pärtel Lippusele partel.lippus@ut.ee

7. Suulise kõne korpus ja dialoogikorpus

Suulise kõne korpuse ja selle ühe allosa - dialoogikorpuse - koostajaks on TÜ suulise kõne uurimisrühm. Suulise kõne korpuse koostamist alustati aastal 1997. See sisaldab suulise kõne lindistusi ning nende litereeringuid, samuti taustainfot.

Korpuse lühitutvustuse leiab Andriela Rääbise koostatud materjalist.

Dialoogikorpus on loodud eelkõige rakenduslikul eesmärgil - arvutil normaalse dialoogi imiteerimiseks, arvuti õpetamiseks infotelefonis vastama vms. Üks selline katsetusjärgus rakendus on näiteks Margus Treumuthi loodud Teatriagent: http://www.dialoogid.ee/teatriagent/

Suulise kõne korpusest loe lisaks: **Tiit Hennoste 2003**. Suulise eesti keele uurimine: korpus. - Keel ja Kirjandus, 7, 481-500.

Dialoogikorpuse kohta vaata veel: **Tiit Hennoste, Andriela Rääbis 2004.** Dialoogiaktid eesti infodialoogides: tüpoloogia ja analüüs. Tartu Ülikooli Kirjastus.

TÜ Eesti suulise keele korpuse ja dialoogikorpuse tutvustus
Andriela Rääbis

TÜ Eesti suulise keele korpus

http://www.cl.ut.ee/suuline/
Kogutud alates 1997

Suulise kõne uurimisrühm
Tiit Hennoste
Olga Gerassimenko
Riina Kasterpalu
Kirsi Laanesoo
Andriela Rääbis
Krista Strandson

Korpuse üldiseloomustus
avatud korpus
universaalne – ei ole teoreetiliselt ette määratud rangelt eri situatsioonitüüpe
Korpus on liigendatud kolmelt aluselt:
argi- ja institutsionaalne suhtlus
dialoogid ja monoloogid
silmast silma, telefoni- ja meediasuhtlus
loomulikud situatsioonid
põhiliselt audiolindistused, videot vähe

Korpuse suurus jaanuar 2009
umbes 360 tundi salvestusi, mis litereerituna oleks umbes 2 160 000 sõna
2049 transkribeeritud teksti
1 346 664 tekstiüksust (sõna ja pausi)

Tekstide arvu järgi
30% silmast silma vestlused (573)
- argivestlused (184)
- institutsionaalne suhtlus (357)
63% telefonivestlused (1320)
- argivestlused (178)
- institutsionaalne suhtlus (1133)
7% meediasuhtlus (152)

Sõnade arvu järgi
54% silmast silma vestlused (729000)
27% telefonivestlused (368000)
19% meediasuhtlus (249000)

Transkriptsioon
sõnad ja mitmesugused suhtlushäälitsused
suhtlusüksused
pausid
kõne omadused (intonatsioon, venitused, katkestamised, rõhud, valjus jne)
pealerääkimised ja haakumised
transkribeerija kahtlused (nt halvasti kuuldud sõnad)
niisuguste nähtuste kirjeldused, mille kohta puudub transkriptsioonimärk või mida ei taheta transkribeerida, kuid mida on vajalik ära näidata.

Näide
T: mhmh
I: tuli üks ´ristmik s=vaatsime=et ´ei, (0.9) ´ju me vale=´koha=bäl oleme.
(1.7)
T: ja ´kõigepealt sellest=ee sõitsime ka ´mööda siiski.
(.)
I: millest.
(1.0)
T: see ku=me olime säl ´suure maantee peal. (0.7) sõitsime ka ju [´mööda.]
I: @ [noh see] oli ´väike asi. @ ((üleolevalt))

Transkriptsioonimärgid
>.....< kiirendatud lõik
<.....> aeglustatud lõik
..... muust kõnest vaiksem lõik
AHA hääle kõvendamine

: venitus
hehe naer
s(h)õnanaerdes öeldud lõik
$.....$ naerev toon
.hh sissehingamine
hh väljahingamine
{-} {---} {sõna} ebaselgused
(( )) kommentaar

Taustakirjeldus
0. Tehniline info
1. Situatsioon ja olukord
2. Suhtlejad, nende omadused ja omavahelised suhted
3. Ainestik ja teema
4. Tekst ja suhtlus
5. Keel ja keelekasutus
6. Lisa

Korpuse kasutamine
Valdav osa lindistusi ei ole avalikud tekstid ega internetis kasutatavad.
Korpus jaguneb eri piirangutasemega alaosadeks.
Korpuse kasutajad sõlmivad lepingu konfidentsiaalsuse kohta.

Märgendus
Osa suulise keele korpusest (100000 sõna) on morfoloogiliselt märgendatud.
http://www.cl.ut.ee/korpused/morfliides/

Eesti dialoogikorpus EDiC

http://www.cs.ut.ee/~koit/Dialoog/EDiC
Korpus on loodud kahel eesmärgil:
uurida inimestevahelist suhtlust,
modelleerida inimese ja arvuti vahelist suhtlust.

Dialoogikorpus
suulised inimestevahelised dialoogid TÜ eesti suulise keele korpusest (1148)
võlur Ozi meetodil kogutud kirjalikud dialoogid (22)
inimese ja arvuti vahelised dialoogid

Suulised dialoogid
- 1148 dialoogi = 221 000 sõna
- 1017 telefonikõnet (infotelefon, reisibüroo, bussijaam, polikliiniku registratuur, kauplused, taksodispetšer jt)
- 131 silmast silma vestlust (kaubandus, teenindus, reisibüroo, teejuhatamine jt)

Aktide tüpoloogia

I. Naaberpaariaktid
Rituaalid (tervitus, tänamine jne)
Teemavahetus
Partneri algatatud parandused
Kontakti kontroll
Direktiivid (soov, ettepanek, pakkumine jne)
Küsimused
Seisukohavõtud (väide, arvamus jne)
II. Üksikaktid
Rituaalid (kontakteerumine, tutvustamine jne)
Infolisad (täpsustamine, pehmendamine jne)
Vabatahtlikud reaktsioonid (jätkaja, vastuvõtuteade jne)
Parandused (eneseparandus)
Primaarsed üksikaktid (eelteade, lubadus, referaat jne)

Märgendatud dialoogi näide

((kutsung)) | RIE: KUTSUNG |
V: .hh info`telefon= | RIJ: KUTSUNGI VASTUVÕTMINE |
| RY: TUTVUSTUS |
Kersti= | RY: TUTVUSTUS |
tere | RIE: TERVITUS |
H: tere= | RIJ: VASTUTERVITUS |
ma=sooviks `Ark (.) `Tartus. | DIE: SOOV |
V: jaa? | VR: NEUTRAALNE VASTUVÕTUTEADE |
üks=`hetk | DIJ: EDASILÜKKAMINE |
(1.8) .hh autore`gister `on `suletud `esmaspäeviti. | YA: INFO ANDMINE |
H: `on jah= | KYE: VASTUST PAKKUV | | PPE: ÜLEKÜSIMINE |
V: =jah, | KYJ: JAH | | PPJ: LÄBIVIIMINE |

Dialoogikorpuse tööpink
http://lepo.it.da.ut.ee/~treumuth/
Võimaldab valida alamkorpusi, teha analüüse ja päringuid.

8. Eesti murrete korpus

www.murre.ut.ee

Murdekorpusest on hetkel võimalik otsida vaid morfoloogiliselt märgendatud tekstidest, lähiajal peaks lisanduma võimalus otsida ka märgendamata tekstist.

Selle teema juures on pikemalt peatutud morfoloogilisel märgendamisel. Kuna murdetekstidest otsimisel on morfoloogiliselt märgendatud teksti olemasolu eriti oluline (keel on väga varieeruv, ise erinevaid vorme ennustada on kohati võimatu), siis on vajalik tunda morf. märgenduse põhimõtteid. Kodutöö puudutabki ühe lühikese murdeteksti märgendamiskatset.

Korpusest, eriti aga morfolooglisest märgendusest ja selle problemaatikast saab lisaks lugeda veel artiklist **Liina Lindström, Liisi Bakhoff, Mari-Liis Kalvik, Anneliis Klaus, Rutt Läänemets, Mari Mets, Ellen Niit, Karl Pajusalu, Pire Teras, Kristel Uiboaed, Ann Veismann, Eva Velsker**. Sõnaliigituse küsimusi eesti murrete korpuse põhjal. – E. Niit (toim.) Keele ehe. Tartu Ülikooli eesti keele õppetooli toimetised 30. Tartu 2006. 154-167.

Korpuse tutvustus

Murdekorpus koosneb eesti murrete helisalvestistest, litereeringutest soome-ugri foneetilises transkriptsioonis, nende lihtsustatud varantidest (nn lihtsustatud transkriptsioon) ning morfoloogiliselt märgendatud tekstidest. Morfoloogilsielt märgendatud tekstid on loetud andmebaasi, millel on veebipõhine otsimootor: http://www.murre.ut.ee/triip/murdekorpuse-otsing/ Lisaks eelnevale on murdekorpuse osaks andmebaas, mis sisaldab infot kõnelejate, salvestiste jms kohta.

Järgnevalt vaatleme murdekorpuse osi ükshaaval.

1**. Helisalvestised**. Helisalvestised on pärit EKI arhiivist ja Tartu Ülikooli eesti murrete ja sugulaskeelte arhiivist. Valdav osa salvestistest on tehtud 1960-1970ndatel. Kõige esimesed salvestised on tehtud 1938. aastal. Kõnelejad on üldjuhul vanad inimesed, keda on intervjueeritud nende kodus. Rollijaotus on väga selge: küsitleja küsib, keelejuht vastab. See on jätnud jälje ka materjalile - korpuse põhjal on raske uurida näiteks küsilauseid murretes. Salvestised on digitaliseeritud.

Lühike helisalvestise näide Rõngu kihelkonnast

2. Litereering foneetilises transkriptsioonis. Helisalvestised on litereeritud foneetilises transkriptsioonis. Kasutatud on klassikalist soome-ugri foneetilist transkriptsiooni. Kuna selles transkriptsioonis on väga palju märke ja märkide kompinatsioone, on praegu foneetilises transkriptsioonis tekst võimalik kasutada vaid programmiga Word koos spetsiaalsete fontidega või pdf-ina.

Litereeringus on taotletud täpsust, st ebakonarused, kordused, partikli, üneemid on samuti litereeritud. Lisaks keelejuhi tekstile on foneetilises transkriptsioonis ka keelejuhi tekst. Järgnevalt on lühike näide samast Rõngu tekstist, mille helifaili ennist kuulsite.

Näide foneetilises transkriptsioonis tekstist

Tartu murre, Rõngu, Pühaste küla, EMH 342
Juuli Antsik (82a, s 1879)
Lind. 1961 H. Keem
Litereering EKIst, üle vaadanud Liina Lindström 25. mai 1999.

JA: *inèmine om̀ ͡jo väèga | näottu ku ͡da | vanaš lät̄ är*
HK: *eìj ͡olè*
JA: *(---) ‖ nigù ärä̀ kujùnu | ťsû̆g ‖*
HK: ((naerab)) *jaȁ | ni ͡eȁt | sā ͡nüt̀ kaš ͡sul ͡om̀ mêlen midä̀gi omà ͡ni laťseʁeľvitsest väřgist nī ͡kaȁ ‖*
JA: *meš ͡sa sā̄lt̀ meĕštaȁ mäleťtädä vī̆l ‖*

Litereeritud tekste on 2009. aasta septembrikuu seisuga u 1 050 000 tekstisõna, 2009. aasta lõpuks peaks mahtu kasvama 1,1 miljonini. Lisaks eesti murrete materjalidele lisandub ka teiste läänemeresoome keelte materjali, eelkõige nende keelte materjali, mis on eesti keelega rohkem kontaktis olnud. 2009. aasta lõpuks peaks lisanduma näiteks liivi keele materjale, plaanis on lisada kindlasti vadja ja isuri materjale.

3. Lihtsustatud transkriptsioon. Lihtsustatud transkriptsioon on vajalik selleks, et mööda minna fon. transkriptsiooni keerukusest ning kasutada tekste ka muude programmidega. Konvertimine foneetilisest lihtsustatud transkriptsiooni toimub automaatselt (st teisendused tehakse automaatselt), käsitsi lisatud on välte märk. Nimekiri lihtsustatud trankriptsioonis kasutatud märkidest:
`kalla - graavis sõna ees märgib, et sõna on 3. vältes
*kalla - tärn sõna ees märgib, et tegemist on 2. ja 3. välte vahelise pikkusega
kal'li - sirge ülakoma tähistab palatalisatsiooni
^ng või ~ng tähistab velariseeritud nn-i (ka^ngas häädub nagu kannas, vrd soome k)
(.) - lühike paus
(...) - pikem paus
= kokkuhäädus
+ liitsõnapiir (ka mõnede produktiivsete sõnataoliste liidete puhul, nt pere+kond)
<com> kommentaari algus (st mitte päris tekst, vaid litereerija kommentaar olukorra vms kohta)
</com> kommentaari lõpp
<u who=KJ> keelejuhi kõnevooru algus
</u> kõnevooru lõpp
<u who=AK> kellegi teise kõnevooru algus, initsiaalid on lahti seletatud faili algul kommentaarides

Rõngu tekstinäide lihtsustatud transkriptsioonis:

<com> Tartu murre, Rõngu, Pühaste küla, EMH 342, KJ = Juuli Antsik (82a, s 1879). Lind. 1961. HK = Hella Keem. </com>

<u who=KJ> inemine omm=jo `väega (.) näottu ku=ta (.) vanass lätt är </u> <u who=HK> eij=ole </u> <u who=KJ> (---) (...) nigu ärä kujunu (.) t'suug (...) </u> <u who=HK> <com> naerab </com> jahh (.) ni=et (.) saa=nüt kas=sul=omm `meelen midägi oma=ni latsõ+bõlvitsest värgist nii=kahh (...) </u> <u who=KJ> mes=sa säält mõistat mälettädä viil (...) </u>

4. Morfoloogiliselt märgendatud tekstid.

Kasutatud on XML-keelt. Märgendamiseks kasutatakse abiprogrammi Mark.

Iga tekstisõna puhul on märgendatud järgmine info:

1. sõne originaalkujul, nii nagu see tekstis esineb: <sne> t's'ibõrdõl'l'i </sne>

2. märksõna kirjakeelestatud kujul: <msn> *t*siberdelema </msn> Kasutatud on kirjakeele ortograafiat, kaotatud on vokaalharmoonia. Kui kirjakeeles on sama tüvega ja sama tähendusega sõna olemas, on märksõnana esitatud kirjakeelne sõna, nt *vaene*, *seal*.

3. tähendus, kui see erineb kirjakeelest: <tah> siplema </tah>

4. morf. vormi kirjeldus ja sõnaklass: <mrf slk="V"ps ind ipf pl 3</mrf>

Vt sõnaklasside loendit ja muutvormide tabelit. Siin ei ole kasutatud klassikalist eesti keele sõnaklasside jaotust (EKG, "Eesti keele käsiraamat", lisatud on mõningaid suulise kõne erijooni.

-info on XML-failis struktureeritud nii, et ühe sõne kirjeldus on märgendite <mark> ja </mark> vahel kindlas järjekorras

5. fraas: <fra>ei õld juo neist *asjagi</fra> Märgitakse peamiselt kinnistunud väljendite puhul.

Näide Rõngu tekstist (ainult algus):

<?xml version="1.0" encoding="ISO-8859-1"?> <!DOCTYPE record SYSTEM "morfo.dtd"> <record khk="RÕN" kla="Pühaste"> <com> Tartu murre, Rõngu, Pühaste küla, EMH 342, KJ = Juuli Antsik (82a, s 1879). Lind. 1961. HK = Hella Keem. </com>

<u who="KJ"><mark><sne>inemine</sne><msn>inimene</msn><mrf slk="S">sg n</mrf></mark> <mark><sne>omm</sne><msn>olema</msn><mrf slk="V">ps ind pr sg 3</mrf></mark>= <mark><sne>jo</sne><msn>ju</msn><mrf slk="Par"/></mark>

<mark><sne>`väega</sne><msn>väga</msn><mrf slk="Adv"/></mark>(.)
<mark><sne>näottu</sne><msn>näotu</msn><tah>kole</tah><mrf slk="A">sg n</mrf></mark>

Lisaks eesti murretele on morfoloogiliselt juba märgendatud ka üle 23000 sõna vadja keelest.

Otsimootori kasutamine

Otsimootor on hetkel olemas vaid morfoloogiliselt märgendatud tekstist otsimiseks ning paikneb aadressil http://www.murre.ut.ee/otsing/search.php

Korpusest saab otsidas sõne, märksõna, tähenduse, sõnaklassi ja muutvormi järgi. Alati peab olema valitud, mis murdest otsitakse (väli *Kihelkond*). Abiks on rippmenüüd. (NB! Kuna otsimootor pole veel päris valmis, ei taha sõnaklassi ja morf. info rippmenüüd hästi töötada, kui nad ei tööta, sisestage vastav otsisõna/lühend käsitsi.)

Muutvormi järgi otsides tuleb silmas pidada, et saab otsida seda, mis sisaldub vormi kirjelduses (nt ind), kui tahta mingit kindlat vormi, tuleb see täielikult kirjeldada, näiteks ps ind pr sg 1 (personaali ehk aktiivi indikatiivi oleviku ainsuse esimene isik). Päritavale vormidele lisaks saab tellida ka konteksti, sel juhul pisut kannatust, sest päringu sooritamine võtab siis rohkem aega.

Väljundit saab ka pisut reguleerida, kui lisada või võtta ära linnukesi kastidest:

Näidata: sõnet☐ märksõna☑ tähendust☐ fraasi☐ sõnaklassi☑ morfi☑ kihelkonda☑ faili☐

Otsimootor ei ole veel lõplikult valmis, siin võib toimuda veel muutusi.

Märgendusjuhend

Morfoloogilisel märgendamisel on abivahendiks programm Mark. Programmi on loonud Karlis Goba. Programm kasutab sisendina txt-formaadis tekste, väljund on xml-is. Programm aitab vältida märgendamisel kergesti tekkida võivaid näpuvigu ning hõlbustab sagedaste sõnade märgendamist.

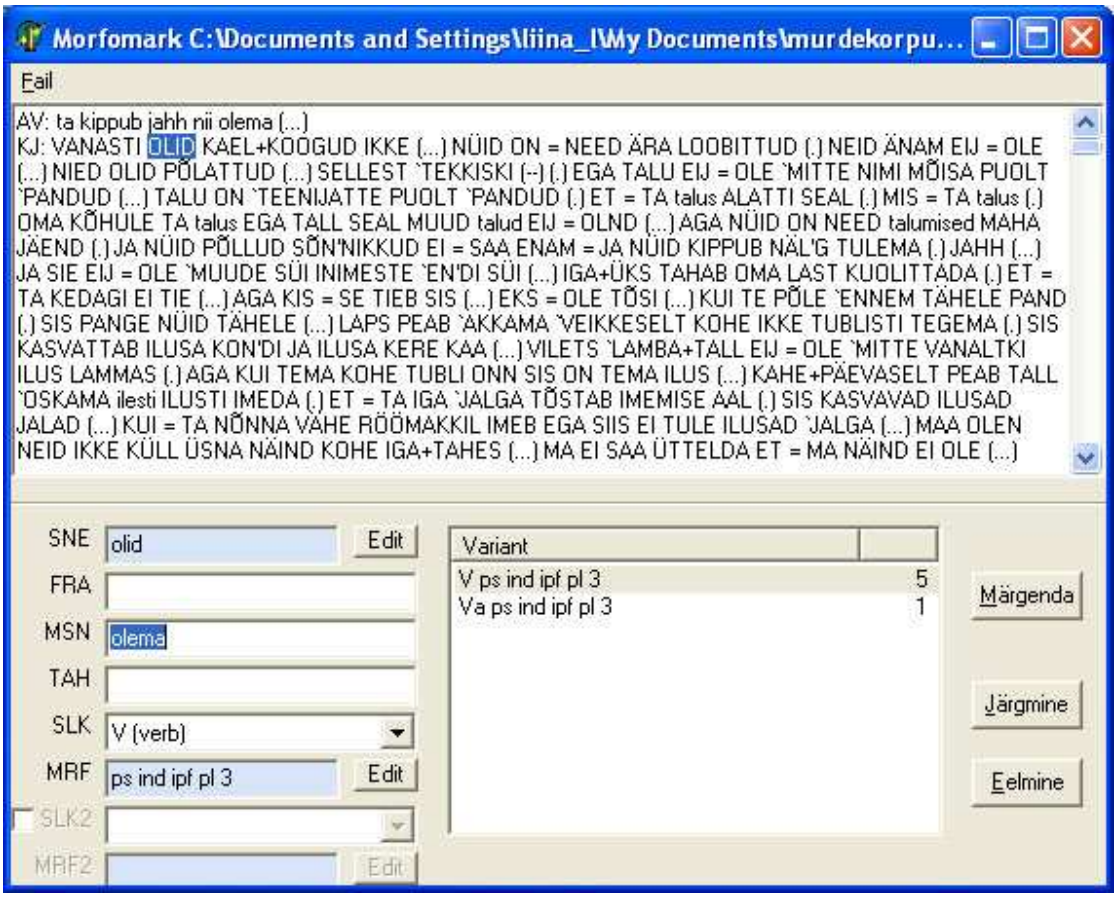
Programmi käivitamine

Tekita arvutisse kataloog morfo vms, laadi sinna failid mark.exe ja morfo.dtd. NB! Need kaks faili peavad kindlasti samas kataloogis paiknema, muidu programm Mark ei hakka tööle.

Käivita programm mark.exe.

Ava fail, mida hakkad märgendama (esimesel korral laiendiga .txt, edaspidi teeb programm sellest faili, mille laiendiks on .xml). Lisa kohe vastavatesse akendesse andmed kihelkonna kohta (suurtähtlühend, nt AVI) ja küla kohta.

Järgnevalt on näide ühe Ambla teksti märgendusest. Juba märgendatud sõnad on suurte tähtedega, märgendamata sõnad väikestega. Märgendada saab ainult keelejuhi (KJ) teksti, käsitleja tekst (näites AV) jääb märgendamata.



Märgendamine

Välja SNE on sõne nii, nagu ta tekstis on.

Iga sõna puhul tuleb kindlasti täita vähemalt väljad MSN ja SLK:

MSN = märksõna kirjakeelestatud kujul, täpsemalt märksõna nii, nagu see kirjakeeles on (vt nt ÕS-i) või "Väikeses murdesõnastikus". Kui sõna on kirjakeeles tundmatu, tuleb nuputada talle ise algvorm kirjakeele ortograafiat arvestades (vokaalharmoonia välja jätta jne). Verbid *ma*-infinitiivis, noomenid ainsuse nominatiivis.

SLK=sõnaklass

FRA= fraas, täita kindlasti ühend- ja väljendverbide puhul (kõigi komponentide märgendamisel). Täita ka muude fraasistunud (=tähenduslikku tervikut moodustavate) elementide ühes lauses kooseksiteerimise korral. Tasub teha *copy-paste* meetodil, et vigu vältida.

TAH = tähendus, täita siis, kui erineb sõna tähendusest kirjakeeles või kui kirjakeeles sõna puudub. Kui tähendust ei oska anda, võib jätta lisamata.

MRF = morfoloogiline info, täita pöörd- ja käändsõnade puhul. Soovitav on kasutada edit-nupu all pakutavaid loendeid, ent võimalik on ka käsitsi täiendada.

Väljal "variant" pakutakse sama sõnavormi varasemaid (s.t samas failis varem tehtud) märgendusi, kiirendab märgendamist (ei ole vaja iga kord vormi eraldi analüüsida). Number sõna taga näitab, mitu korda seda sõnavormi on selles tekstis juba esinenud.

Nupud "eelmine" ja "järgmine" aitavad algfailis liikuda.

Kui ei oska/taha mingit sõna kohe märgendada, võib selle jätta esialgu märgendamata ja teha seda hiljem (liigu edasi nupu abil "järgmine"). Märgendamata sõnad on väikses kirjas, juba märgendatud sõnad suurte tähtedega.

Liitsõnu, mis on tähistatud +-ga, käsitletakse ühe sõnana ja neid märgendatakse koos. Kui liitsõna on millegipärast algfailis teisiti märgitud (nt kokkuhäädusega), saab algfaili parandada nupu all edit (välja SNE juures). Ettevaatust! Juba märgendatud teksti ei tohi edit-aknast parandada, hiljem võib juhtuda, et fail ei avane. Kui olete liitsõna ühe osa juba märgendanud enne edit-aknas parandamist, jätke järgnev sõnaosa (st liitsõna teine pool) parem märgendamata. Seda saab parandada hiljem mõne tekstitöötlusprogrammiga (Notepad, Wordpad), aga sel juhul peab olema ülimalt ettevaatlik, et xml-koodi mitte ära rikkuda.

Poolelijäänud sõnad märksõna ei saa, sõnaklassiks pange neile X.

Järgnev osa on ülevaade märgendamisel kasutatavatest sõnaklassidest, tunnustest ja lõppudest.

Morfoloogilisel märgendamisel eristatud sõnaklassid ja nende lühendid.

Substantiiv (nimisõna)		S	nt *kas'*
Pärisnimi		H	*Jüri, Pä rnumaa*
Verb (tegusõna)		V	nt *ostma*, v.a *olema* abiverbina
	Abiver b	Va	ainult *olema*-verbi vormid
Adverb (määrsõna)		Adv	nt *täna*
	Proadverb	ProAdv	*siin, seal, kunagi, kõikjal*
	Abimäärsõna	Adva	verbi osad, nt *välja* (*mõtlema)*

Numeraal (arvsõna)			
	põhiarvsõnad	Nump	nt *kaks*
	järgarvsõnad	Numj	nt *teine*
Adjektiiv (omadussõna)		A	*vana*
	Adjektiiv komparatiivis	Ak	*vanem*
	Adjektiiv superlatiivis	As	*vanim*
Pro-sõnad			
	Prosubstantiiv	ProS	Nt *see, too, tema, mina*
	Proadjektiiv	ProA	Nt *niisuke, sihuke*
	Proadverb	ProAdv	Nt *siin, seal*
Kaassõnad			
	Postpositsioon	Post	nt *maja* ***taga***
	Prepositsioon	Pre	nt ***pärast*** *sööki*
Diskursusepartiklid		Par	nt *noh, jah, no, oi*
Suhtlussõnad		Suht	*aitäh, palun, tere*
Onomatopoeetilised sõnad		Ono	*mürts*
Küsisõna		Intr	nt *kas*, ka relatiiv-interrogatiivpronoomen (*kes, kelle, keda* jne), küsivad-siduvad adverbid (*miks, millal, kus* jne)
Konjunktsioon (sidesõna)		Konj	nt *ja, et*
Eitussõna liitvormides		Mn	*ei, mitte*
Võrdlussõna liitsuperlatiivis		Ms	*kõige*
Muud		Muu	ei oska määrata või sobiv kategooria puudub
Määramata		X	ei ole võimalik määrata (poolikud sõnad)

· Adjektiivide puhul tuleb märksõnaks alati märkida sõna algvorm (kompareerimata vorm, nt vana, ilus), ka siis, kui sõnaliigi märgendiks tuleb Ak.

· Kliitikuid (ki/gi-liide) märgitakse nagu liitsõnal plussiga märksõnas, ja märksõnas ainult *ki*-vorm, nt sne *emagi* msn *ema+ki*

· Prosubstantiividel märgitakse märksõnaks pikk vorm (*mina, sina, tema, meie, teie, nemad*), olenemata sellest, kas teistis oli lühike või pikk.

· Proadjektiividel (ProA) tuleb praegu kogu morf. info lisada käsitsi.

· Põhi- ja järgarvsõnad: kui on kirjekeeles mitmest sõnast koosnev arvsõna, siis kirjutame selle kokku liitsõnaks (nii sõne kui märksõna väljal, nt *kaks+kümmend+viis*)

3. Käänduvad sõnad:

noomen, pronoomen, adjektiiv, pärisnimi, numeraal

Tabel 3. Käändsõnadele (S, ProS, A, Ak, As, ProA, H, Nump, Numj, Intr) märgitud morfoloogilised tunnused ja lõpud.

ainsus	sg
mitmus	pl
nominatiiv (nimetav)	n
genitiiv (omastav)	g
partitiiv (osastav)	p
illatiiv (sisseütlev)	ill
additiiv (lühike sisseütlev)	add
inessiiv (seesütlev)	in
elatiiv (seestütlev)	el
allatiiv (alaleütlev)	all
adessiiv (alalütlev)	ad
ablatiiv (alaltütlev)	abl
translatiiv (saav)	tr
terminatiiv (rajav)	ter
essiiv (olev)	es
abessiiv (ilmaütlev)	ab
komitatiiv (kaasaütlev)	kom
instruktiiv (viisiütlev)	inst
possessiivsufiks	poss

· Aditiivi tegelikus elus me ei kasuta, sest 1) ta siiski pigem ei ole omaette kääne; 2) murretes on väga paljudel juhtudel raske otsustada, millal on pikk sisseütlev, millal lühike sisseütlev (aditiiv).

Tabel 4. Pöördsõnadele (V, Va) märgitud morfoloogilised tunnused ja lõpud.

da -infinitiiv	inf
des -gerund	ger
ma -supiin	sup
tav -partitsiip	tav
nud -partitsiip	nud
tud -partitsiip	tud
v -partitsiip	v
Personaal	ps
Impersonaalne passiiv	ips

Personaalne passiiv	pas
Indikatiiv (kindel)	ind
Konditsionaal (tingiv)	knd
Imperatiiv (käskiv)	imp
Jussiiv (möönev)	jus
Kvotatiiv (kaudne)	kvt
Potentsiaal	pot
Preesens (olevik)	pr
Imperfekt (lihtminevik)	ipf
Ainsus	sg
Mitmus	pl
Esimene isik (mina, meie)	1
Teine isik (sina, teie)	2
Kolmas isik (tema, nemad)	3
eituse märgend verbivormi lõpus	neg
eitussõna	Mn

· Märgendatakse eraldi iga sõnavorm. Sellest tulenevalt ei ole võimalik liitaegu (täis- ja enneminevikuvorme) üheskoos täis- ja enneminevikuks märgendada, vaid olema-verb märgendatakse abiverbiks (Va) ja vorm lähtub selle sõna vormist, nud-partitsiip märgendatakse põhiverbi (V) nud-partitsiibiks.

· Personaalset passiivi märgitakse vaid sel juhul, kui tegemist ei ole liitajaga (st eesti keele lause Tööd olid tehtud siia alla ei kuulu, selles märgendatakse eraldi olema-verb (Va) ja tud-partitsiip.

· Kui partitsiibile järgneb käändelõpp või arv, lisatakse see käsitsi lõppu, nt Võru *(sai) tettüss* V tud tr

· Impersonaal: märgendusprogramm lisab automaatselt lõppu ka isiku: nt *söödi* V ips ind ipf sg 1 --> siit tuleb lõpust isik käsitsi ära kustutada, peab olema ips ind ipf.

· Eituse puhul saab eitussõna märgendi ei (mõnedes murretes ka *mitte, es, ep* jne) ning verbi eitusvorm saab oma märgendi neg. Oleme kokku leppinud, et kui eitustüvel isikulõppe ei ole (ja tavaliselt ju ei ole, nt *ei tule*), siis eemaldame ka eitusest isikulõpud käsitsi, st õige märgendus oleks V ps ind pr neg. Eituse minevikuvormis on eitussõna ja *nud*-partitsiip (*ei teinud*); nud-partitsiibil me eitust märkinud ei ole.

KOMMENTAARID, LOENDID

Pro-sõnad			
	Prosubstantiiv	ProS	Nt *see, too*
	Proadjektiiv	ProA	Nt *niisuke, sihuke*

	Proadverb	ProAdv	Nt *siin, seal*

· Praegu pole oma märgendit asearvsõnadel *mitu, mitmes, mitu-setu, mitmes-setmes, mitmendik.* Võib märgendada .

· Prosubstantiividel märgitakse märksõnaks pikk vom (*mina, sina, tema, meie, teie, nemad*), olenemata sellest, kas teistis oli lühike või pikk.

· Atribuudina kasutatavad asesõnad *see (mees), too (naine), üks (vanamees)* on kokkuleppeliselt märgitud prosubstantiivideks.

· sõna *kõik* – märgitakse prosubstantiiviks (ProS) ja arv (ainsus või mitmus) tema sisulise ainsuslikkuse või mitmuslikkuse alusel (kõik mehed – pl; kõik maailm – sg, 'kogu')

Prosubstantiivid: vt EKK 145-...

1) isikulised asesõnad e personaalpronoomenid: *mina,sina tema, meie, teie, nemad*

Võru: *maq, saq, tää ~tiä ~timä, miiq, tiiq, nääq*

eeR, eeK: *minä ~mä ~mina ~ma, sina ~sa ~sinä ~sä, meije ~ mei ~me, teije ~ tei ~te, tämä ~tä ~ta ~tema, nämä(d) ~nämad ~nääd ~ näd ~nemäd ~näväd*

2) enesekohased asesõnad e refleksiivpronoomenid näitavad, et tegevuse objekt langeb kokku tegevuse subjektiga, tegevus on suunatud tegijale endale: *enese ~ enda, iseenese ~ iseenda, oma.*

eeR, eeK: *enese, henes(ä),* taval. koos poss. sufiksitega

3) omastavad asesõnad e possessiivpronoomenid näitavad, et täiendiga vormistuv omaja langeb kokku tegevuse subjektiga: *oma, enese ~ enda, iseenese ~ iseenda, iseoma, omaenese ~ omaenda*

4) vastastikused asesõnad e retsiprookpronoomenid: *üksteise, teineteise.*

5) Näitavad asesõnad e demonstratiivpronoomenid: *too, sama, seesama, toosama, teine, muu,*

Võru*: seo, taa, tuu, neoq, naaq, nuuq*

eeR, eeK: *sie~se* (g. *sene*), *nie~ne~nied*; pejorat. varjundiga: *tuo, nuod*; murrakuti ka *tämä*

6) Määratlevad asesõnad e determinatiivpronoomenid: *ise, oma, iga, igaüks, igamees, kõik, mõlemad, kumbki, emb-kumb, kogu, terve.*

7) Umbmäärased asesõnad e indefiniitpronoomenid: *keegi, miski, ükski, mõni, paljud, üks, teine*

eeR, eeK: *keski, kieki, kiegi, kedagi, kennegi, kengi, kenegi, miski, miˇnnegi, menegi, migägi, midägi, jotaki, jodagi, muu* (g *munde, muije*)

Proadjektiivid : vt EKK147-148: näitavad e dem. pron: *niisugune, samasugune, niisamasugune, selline, seesugune, säärane, säherdune, taoline, selletaoline, nisuke, niuke, nihuke, sihuke, siuke*. Murretes palju variante

umbmäärased: *mingi, mingisugune, miskisugune, mõningane, mõningad*

eeR, eeK: *niisugune, niisukene, niisukaine, niske, nisike, nisune, neske, monikane, mõnikane, monikaine*

Proadverb : *siin, seal, sinna, sealt, siia, siit, siiapoole, siinpool, siitpoolt, siiasamasse, siinsamas, siitsamast jne. siis, nii, nõnda, sedasi, sedaviisi, sedamoodi, sedavõrd, niivõrd, seepärast, sellepärast, seetõttu, seeläbi, mistõttu, kusjuures, seevastu, sellegipoolest, kunagi, millalgi, kusagil, kusagile, kusagilt, millegipärast, miskipärast, mujale, mujal, mujalt, teisale, teisal, tesailt, teisiti, kõikjal, kõikjale, kõikjalt, alati, igati.*

Küsisõna (Intr)

siia hulka on arvatud ka interrogatiiv-relatiivpronoomenid

kas, või, vä, kes, missugune, mis, kumb, milline, mäherdune, misuke, mitu, mitmes, mitmendik, kuhu, kus, kust, millal, kuidas, miks, millepärast, misjaoks, mistarvis ,

Võrus: *kiä~keä, miä~meä, mille* ‘miks’, *kuis* ‘kuidas’, *kuimuudu, määne* jne

eeR, eeK: *kie, ke, kiese, kenn, kense, kens* (g *ken, kene, kenne*), *mie, mi, me, mikä, miga, mige,* (g *mine, miˇnne*), *miˇnnesugune, messugu(ne), missugune, misuke, missukaine, missikene, mitäsugune, mingalaine, milla, millas, kõõs* ‘millal’

Adverb (määrsõna)		Adv	nt *täna¸* sh rõhumäärsõnad
	Abimäärsõna	Adva	verbi osad, nt *välja* (*mõtlema)*

Adverb :

Tavalised adverbid

EKK: määrsõnad on muutumatud sõnad, mis esinevad lauses määrusena. Vt lk 142-144

Rõhumäärsõnad : vt EKK lk 157-158. R. on muutmatud sõnad, mis toimivad lauses üldlaienditena (lause- ja fraasilaienditena). Nad annavad edasi kõike seda, mis käib

lausega väljendatava sündmuse kui terviku kohta: tõenäosus- v modaalhinnanguid, nt *võib-olla, arvatavasti, vist, kahjuks*, suhtluseesmärki (*kas, las*), sündmuse või selle osaliste tuntust, olulisust vms kuulaja jaoks, nt *ju, siis, ka, samuti, veel, eelkõige, hoopis, küll, just, kas või, juba, no, ometi, jah, vaat, lihtsalt.*

Rõhumäärsõnadel oma märgendit ei ole, märgendame need kas määrsõnadeks (pikemad, leksikaalsem sisu, harvemini kasutatavad) või Diskursuspartikliteks (Par; lühemad, sagedasemad).

Abimäärsõna : vt EKK, lk 151-152. Abimäärsõnad on muutumatud sõnad, mis kuuluvad lauses tegusõna juurde, andes sellele mingi uue tähendusvarjundi või konkretiseerides tegusõna tähendust. Moodustavad koos tegusõnaga uue tähendusliku terviku. nt *läbi (elama), vastu (võtma), kallale (kippuma), tagasi (tõmbuma), ära (sõitma), ümber (aelema), valmis (saama), laiali (valguma), läbi (paistma), sisse (elama), ära (minema), läbi (lugema, elama, vaatama).*

Sidesõnad

ja, ning, ega, ehk, või, aga, kuid, ent, vaid, et, kui, kuna, sest, kuni, kuigi, ehkki, nagu, saati, elik, justkui, otsekui

murretes ja/või suulistes tekstides: *a, aq, ni(q)* (‘ja’, Võru), *ku, t* ‘et’

Kaassõnad

Muutumatud sõnad, mis kuuluvad lauses nimisõna juurde, andes sellele ligilähedaselt samasuguseid tähendusi nagu käändetunnused, nt (maja) taga, (linna) kohal, (laua) ümber, (saali) keskel, mööda (teed), pärast (sööki), enne (vihma).

Kaassõnad			
	Postpositsioon	Post	nt *maja **taga***
	Prepositsioon	Pre	nt ***pärast** sööki*
Võrdlussõna liitsuperlatiivis		Ms	*kõige*

eeR; eeK: *kaige ~kaikse, kõige ~ kõikse*

Possessiivsufiks	poss

lisandub arv ja isik, nt poss sg 3

eeR, eeK: sg 1 –ni, -ne

pl 1: -mme

sg 3: -sa, -sä, -se, -s

Suhtlussõnad

- *tere, tervist, tsau, nägemist, nägudeni, head aega, teie soov palun* jms; *jõudu tööle, head isu, jätku leivale, terviseks, palju õnne*; *saage tuttavaks, saame tuttavaks*; *palun, võtke heaks, tänan, aitäh*; *vabandust, andke andeks, pole midagi*; *läks, start, stopp*.

- *siga, kurat, raisk*; *kaabakas, mölakas, loll; õudne, lõpp, jama, jumal hoidku, no kuule, mine põrgu*; *issand, jessas*; *just, justament, ausõna, selge, hästi*.

Partiklid (suulise kõne korpuse põhjal)

- *ahah, ahaa, aa, ah, aah, ahhaa*; *mhmh, mhm, mhõh, mh, mm, nii, mhh, mmh, mmm*; *jaajaa, jaajah*.

-hm, aa, ah, ohhoo, ohoh, oih, oi, oo, assa, ossa, vau; *ah, oo, mm*; *hurraa, haa, vau*; *oohh, uh*; *hurraa, jess, jee, ah, oo*; *voh*; *einoh, jah, jess.*

- *nonoh*; *ah, oh*; *ah, eh, oeh*; *oih, oh, oi*; *aaa*; *ai, aih, aii, ah*; *häh, päh, fui, fuh*; *öäk, pthüi*; *ahhaa*; *äähh*; *hmm, tjah, mnjaa* / *mnjah, nooh*; *hõh, näh, ah*.

- *khõm, hm, halloo, uu, hei, ahoi, uhuu, ts*; *uu, halloo, hei, ahoi*; *ptruu, ass, nõõ, võts, kis-kis, miki-miki*; *kuss, pst, ts*; *säh, näh, noh*; *kõtt, kõss, hurjuh*; *oot-oot, nonoh*.

- *ee, õõ, öö, aa*; *noh, nagu, see*; *või, või ühesõnaga, või siis, või parem, tändab/tähendab, ei/mitte.*

- *mhmh/mh, odot*; *vä /võ/või*; *jah/jaa, ahah/ah, et, eks*; *ei, tähendab/tähendap, ikka*.

- *et* + eelmise vooru kordus; *jah, jajah, jaa, tjah, jaah, ja=jaa, mhmh, mqm*; *ei, äqä, mitte + aga.*

9.-10. Unixi/Linuxi töövahendid korpustega tegelejale

Seni on sellel kursusel tutvustatud peamiselt veebipõhiseid otsimootoreid, mis nõuavad kasutajalt küllalt vähe eelteadmisi. Eelteadmisi on vaja peamiaselt selle kohta, mis korpuses on ning mida sealt oodata. Selliste veebipõhiste otsimootorite peamine puudus on, et nende otsingu sooritamise võimalused on piiratud, need on ette antud otsimootri koostajate poolt. Samas pakub iga korpus võimalusi märksa mitmekesisemateks otsinguteks ja korpusematerjalide töötluseks. Selleks pakub häid tingimusi Unixi/Linuxi operatsioonisüsteem, mis on kasutusel enamasti suurtes serverites, aga paljudel inimestele ka oma koduarvutis. Unixis/Linuxis on palju lihtsaid käske, mis võimaldavad teksti töödelda väga erineval moel. See ja järgnevad teemad ongi pühendatud nende töövahendite tutvustamisele. Me ei jõua selle kursuse raames nende töövahendite tuvustamisega väga sügavale minna, ent alati on võimalik ise edasi õppida, kui algus käes on.

Selles osas tutvustame Unixi/Linuxi käske, mis on seotud failide haldamise ja kataloogides liikumisega. See on alus edaspidisele.

Unixi kohta on internetis palju erineviad kasutusjuhendeid, näiteks http://www.msg.ucsf.edu/local/programs/ono/manuals/unix_for_beginners.html

Enamik neist siiski pole mõeldud päris lingvistile, seepärast olen järgnevatesse teemadesse koondanud just keeleteadlastele vajaminevate käskude juhendeid.

Töö toimub ülikooli serveris adalberg.ut.ee, kuhu sisenemiseks on teil vaja oma kasutajanime ja parooli. Need on samad, mida kasutate ülikooli meili lugemisel või näiteks ÕISi sisenemisel.

Serverisse sisenemiseks on soovitav kasutada programmi SSH Secure Shell Client või Putty. Programmid ja kasutusjuhendid leiate ülikooli IT-osakonna leheküljelt: http://www.ut.ee/4180

SSH Secure Shell Client tuleb installida oma arvutisse. NB! Kui endal installimine ei õnnestu, on võimalik, et teil ei ole oma arvutis administraatori õigusi (õigusi installeerida uusi programme). Pöörduge kellegi poole, kellel need õigused on.

Puttyt saab kasutada ka ülikooli veebilehe kaudu. Jälgige IT-osakonna leheküljel http://www.ut.ee/4180 paiknevaid juhendeid.

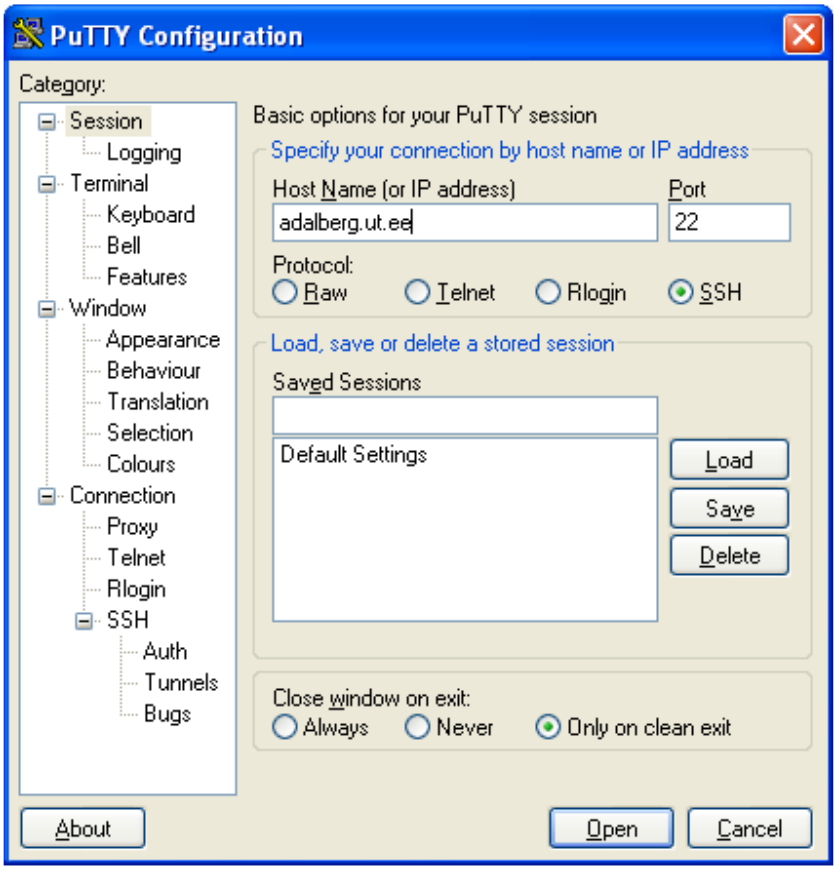
Kursuse materjal koosneb käskude selgitustest, mis tuleb kõik kohe praktiliselt läbi teha, et toimimispõhimõtetest aru saada.

UNIXi/LINUXi kasutusjuhend keeleteadlastele

Sisselogimine

Masinasse adalberg.ut.ee sisselogimiseks kasuta programmi SSH Secure Shell Client või Putty.

Puttyt kasutades:
Puttyt saab kasutada otse ülikooli it-osakonna veebilehelt, ei nõua eraldi installimist:
http://www.ut.ee/4324
Sisesta masina nimi (host name): adalberg.ut.ee
vali protokolliks SSH



Seejärel küsitakse kasutajanime ja parooli. Need on samad, mis sul ülikooli võrgus üldiselt (nt ÕISi logides).

SSH-d kasutades:
ka programmi SSH Secure Shell Client leiate ülikooli IT-osakonna veebilehelt. See tuleb installida oma masinasse. Kasutamiseks:
- vajuta Quick Connect

- sisesta masina nimi (host name): adalberg.ut.ee
- oma kasutajanimi (User name), see on kasutajanimi ülikooli arvutivõrgus (sama, mis nt ÕISi sisenemisel)
- vajuta Connect.
Seejärel küsitakse sult parooli (sama, mis ÕISi sisenemisel).

Kui lõpetate töö ja lahkute arvuti juurest, tuleb end kindlasti ka serverist välja logida. Selleks kirjutage käsureale exit.

Kodeeringust

Täpitähtede jms nägemine ja sisestamine sõltub masinas kasutatavast kodeeringust. Tänapäeval on kõige tavalisemaks kodeeringuks, mida masinad kasutavad, UTF-8, see on kasutusel ka Tartu ülikooli serverites. Puttyga sisselogimise järel näetigi õpetust, kuidas masinasse sisselogimise järel oma Putty terminaliaken seadistada nii, et see näitaks serveri pilti korrektselt. Oletuslikult kasutab Putty kodeeringut ISO-8859-1, mis on varasem standard. Ka ssh ei ole seadistatud UTF-8 peale, nii et täpitähtede sisestamise ja nägemise puhul võib olla probleeme.

UTF-8 on ilmselt ka korpuste tulevik, ent praegusel hetkel on vähemalt murdekorpuses kasutatud varasem kodeeringustandard ISO-885915. Selleks, et server, kasutajaprogramm Putty või SSH ja korpused omavahel kenasti läbi saaks, on mõttekas kasutada seega ISO standardit. Selleks tuleks sisselogimise järel anda masinale käsk
setenv LC_ALL et_EE.iso885915

Selle käsu sooritamise järel kasutab ka server kodeeringut ISO-8859-15 ja peaks korrektselt aru saama sisestatud täpitähtedest ja korpuses kasutatud täpitähtedest.

Seadistus kehtib kuni seansi lõpuni, st kuni väljalogimiseni.

Töö failidega
Unixi/Linuxi shelli kasutades unusta ära, et arvuti küljes on hiir, sellega pole siin midagi teha! Käske saab anda ainult neid kirjutades. Selleks peab neid aga teadma. Siin dokumendis on toodud käsud, mida on vaja lihtsaks tööks tekstifailidega.

Rohkem infot iga käsu kohta:
käsk --h (näiteks cd –h)

Põhjalikum juhend: man käsk (nt man uniq)
Unixi kohta on internetis väga palju kasutusjuhendeid, näiteks
http://www.msg.ucsf.edu/local/programs/ono/manuals/unix_for_beginners.html
Enamik neist siiski pole mõeldud päris lingvistile, seepärast olen siia koondanud just keeleteadlastele vajaminevate käskude juhendeid.

Liikumine kataloogides, kataloogi sisu vaatamine

Unixis /Linuxis paiknevad failid kataloogides nagu muudeski operatsioonisüsteemides. Kataloogipuud saavad alguse nö juurikast. Näiteks paikneb murdekorpuse väike õppekorpus kataloogipuus järgmiselt:
/home/murakas/l/liina_l/murdekorpus
/ kõige ees tähendab, et kataloogipuud jälgitakse nö juurest alates (absoluutne tee)
murakas on kataloog, milles on väga suur hulk ülikooli kasutajate kodukatalooge
liina_l on Liina Lindströmi isiklik kodukataloog, normaalselt (kui ma pole lubanud teisiti) ei ole teistel sellele ligipääsu.
Murdekorpus on Liina Lindströmi kodukataloogis olev kataloog, millele on selle kursuse jaoks antud ligipääsuõiguse ka teistele inimestele.

Seda, kus kataloogis sa parasjagu paikned, näeb käsuga pwd.
Kataloogide vahel liikumiseks on käsk cd.

See, kus kataloogis parasjagu oled, ilmneb ka käsurealt (näites punasega märgitud). Järgnevas näites ollakse liina_l kodukataloogi alamkataloogi kirjkorpus alamkataloogis 1930_ilu_ttxt.

[125] liina_l@adalberg:~/kirjkorpus/1930_ilu_ttxt>

Siia taha kirjutatakse käsud.

Harjuta:
logi end sisse masinasse adalberg.ut.ee
selgita välja, kus paikneb sinu kataloog (käsk pwd; väga oluline edaspidi teada!)

mkdir nimi	kataloogi *nimi* loomine
cd nimi	kataloogi vahetamine (*change directory*), liigud kataloogi *nimi*
cd ..	liigud kataloogipuus ühe sammu võrra tagasi
cd	liigud oma kodukataloogi
pwd	näitab, kus kataloogis oled
ls	sirvib faile (lühike versioon)
ls -l	pikk versioon, annab järgmist infot:
	õigused omanik gruppsuurus muutmisaeg nimi

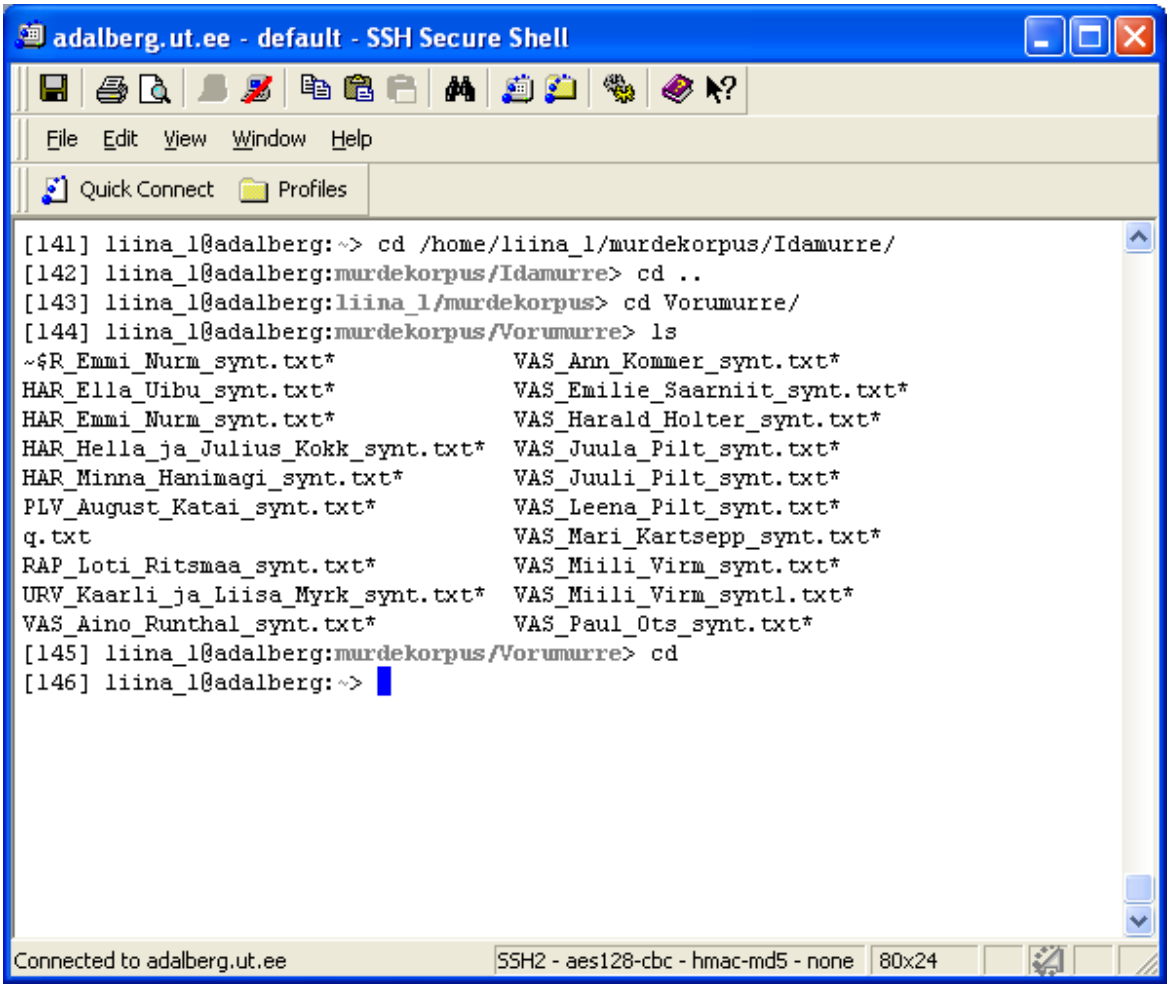
Vihje:
Failinimede puhul on võimalik osa tekstist asendada *-ga. * tähistab üht või mitut sümbolit, mis pole määratletud, st märgib, et failinimes võib selles osas olla misiganes. See on eriti kasulik, kui ühe käsuga tahetakse manipuleerida mitme failiga. Kui tahetakse liigutada, kopeerida, kustutada ainult üht faili, siis on mõistlikum kogu failinimi välja kirjutada.

Näiteks käsuga cp *.txt murre
kirjutatakse kõik .txt-lõpulised failid kataloogi murre.

mv KOD* murre liigutatakse kõik failid, mille nimi algab KOD-iga, kataloogi murre.

Murdekorpuses on failinimed üldjuhul organiseeritud nii, et failinime alguses on suurte tähtedega lühend kihelkonnast, millest see tekst pärit on.

Järgmisel ekraanipildil on näidatud, real [141] kuidas kõigepealt liigutakse kataloogi murdekorpus alamkataloogi Idamurre; [142] sealt sammu võrra tagasi (kataloog murdekorpus); [143] sama kataloogi alamkataloogi Vorumurre; [144] vaadatakse kataloogi Vorumurre sisu; [145] liigutakse oma kodukataloogi.



Järgmises näites on kasutatud pikemat failide loendamise käsku ls –l, mis annab lisaks kataloogis olevate failide ja alamkataloogide nimedele ka infot õiguste, omaniku, grupi, faili suuruse ja viiamse muutmisaja kohta.

Näide:

```
[103] liina_l@adalberg:~/murdekorpus> ls -l
total 64
õigused      omaja    grupp    faili suurus  viimane muutmisaeg   faili/kataloogi nimi

drwxr-xr--   2 liina_l  users       512 dets  3  2007 bin/
drwxr-xr--   2 liina_l  users       512 dets  3  2007 Idamurre/
drwxr-xr--   2 liina_l  users      1536 dets  5  2007 Saartemurre/
drwxr-xr--   2 liina_l  users      1024 dets  5  2007 Vorumurre/
Kataloogis Idamurre:
-rwxr--r--   1 liina_l  users      5855 dets  3  2007 KOD_Anna_Lindvere_synt.txt
-rwxr--r--   1 liina_l  users      6210 dets  3  2007 KOD_Jakob_Luka_synt.txt
-rwxr--r--   1 liina_l  users     20039 dets  3  2007 KOD_Miili_Lepp_synt.txt
-rwxr--r--   1 liina_l  users     21166 dets  3  2007 KOD_Miili_Mardijarv_synt.txt
```

Õigused

Õigused faili vaadata, kirjutada ja käivitada on kirjeldatud kolme rühma kohta: 1) faili omaja (näites liina_l), grupp (näites users), ülejäänud masina kasutajad. Õigused on kirjeldatud 10-kohalise koodina.

Selles 1. koht näitab, kas tegemist on kataloogiga või mitte:

```
drwxr-xr--   2 liina_l  users       512 dets  3  2007 bin/      on kataloog
-rwxr--r--   1 liina_l  users      5855 dets  3  2007 KOD_Anna_Lindvere_synt.txt ei ole
```

kataloog, vaid on fail

Järgmised kolm kohta kirjeldavad faili omaniku õigusi:

r – faili lugemisõigus (*read*)
w – faili kirjutamisõigus (*write*)
x – faili käivitamisõigus (*execute*)

Järgmised kolm kohta kirjeldavad grupi õigusi, samuti r, w, x. Et teada saada, mis gruppi sa kuulud, küsi oma gruppe nii:

groups sinu_kasutajanimi

Näide:

```
[105] liina_l@adalberg:~/murdekorpus> groups liina_l
users
```

See tähendab, et kasutaja liina_l kuulub gruppi users. (Sinna kuuluvad kõik kasutajad, osal kasutajatel on muidugi rohkem gruppe.)

Harjuta:
liigu oma kodukataloogi ja vaata, mis selle sees on: ls
tee sinna uus kataloog, mille nimi on murre: mkdir murre
Liigu oma kodukataloogist õppekorpusesse (siina antakse ette tee nö failipuu päris juurelt alates):
cd /home/liina_l/murdekorpus
liigu murdekorpuse sees Idamurdesse: cd Idamurre
liigu sealt tagasi: cd ..
liigu Saarte murde kataloogi: cd Saartemurre
vaata selle sisu: ls
kopeeri sealt kõik txt-laiendiga failid oma kodukataloogi alamkataloogi murre: cp *.txt /home/sinu_kasutajanimi/murre/
liigu oma kodukataloogi alamkataloogi murre ja vaata, kas failid on seal kenasti olemas, vaata ka failide õigusi:
cd
cd murre
ls –l

Vihje: käsuread on sageli pikad ja tüütud iga kord kirjutada, eriti kui näiteks käsurida läheb vussi ja tuleb otsast alata. Siis oleks mugavam võtta eelmine käsurida ja seda modifitseerida. Selleks mõned võtted:

nool üles: eelmine käsk, mille sisestasid/sooritasid
Cntrl-u käsurea kustutamine (vajuta Cntr-klahv ja u-klahv korraga alla!)
Cntrl-a ***käsurea algusesse***

Failide kopeerimine, nime muutmine, kustutamine

cp fail1 fail2 faili kopeerimine failist1 faili2
cp fail /home/kasutajanimi
faili kopeerimine oma kodukataloogi (kasutajanimi = sinu kasutajanimi)
mv fail1 fail2 failinime muutmine, faili ümbertõstmine (failist1 faili2)
rm fail faili kustutamine

Harjuta:
- Loo oma kodukataloogis kataloog murdekorpus. mkdir murdekorpus
- Kopeeri sinna üks fail Idamurdest (selleks liigu cd-ga eelnevalt kataloogi /home/murakas/liina_l/murdekorpus)

    cp KOD_Anna_Lindvere_synt.txt /home/sinu_kodukataloog/murdekorpus
- liigu sinna kataloogi: cd /home/sinu_kodukataloog/murdekorpus

(või 2 käsku:

    cd

    cd murdekorpus)
- nimeta see fail ümber failiks proov.txt

    mv KOD_Anna_Lindvere_synt.txt proov.txt

kustuta see fail

    rm proov.txt

Vihje: pikkade failinimede ümberkirjutamine iga kord on väga tüütu ja lihtne on teha näpuvigu. Selle tegevuse hõlbustamiseks on TAB-klahv (kust tavaliselt saab taanet): sisesta failinime (või ka kataloogi) esimesed tähed ja vajuta TAB-i, siis pakutakse sulle failinime/kataloogi nime lõppu.

Tekstide sirvimine

Kõige lihtsam tekstieditor pico, mõnes serveris on kasutusel ka analoogiline nano.

Kui fail on olemas:
pico failinimi faili loomine või editeerimiseks avamine
(või: nano failinimi)
Kui faili pole, luuakse sama käsuga uus fail.
Programmist pico või nano väljumine: Cntr-x, akna allservas ka muud valikuvariandid.
Kui failis on tehtud muutusi, küsitakse väljumisel, kas salvestada.

Faile saab avada lugemiseks/vaatamiseks/uute käskude rakendamiseks ka teisiti, näiteks:

cat fail	avab terve faili (suunab standardväljundisse, st ekraanile)
tail fail	avab faili lõpu (oletuslikult 10 rida)
tail -25 fail	avab viimased 25 rida failist
head fail	avab faili alguse (oletuslikult 10 rida)
head -25 fail	avab esimesed 25 rida failist
more fail	võimaldab faili lehekülghaaval lehitseda (saab lehitseda ainult edaspidi) siin edasiliikumine tühikuklahviga katkestamine Cntrl-c

Kui soovid avada/sirvida korraga mitut faili, siis tuleb kasutada failinimede ühisosa (näiteks seda, et failid algavad või lõpevad ühtmoodi). * tähistab mistahes arvu sümboleid:

cat KOD_*	avab kõik failid, mille alguses KOD_
cat *.txt	avab kõik .txt-laiendiga failid

| "toru" võimaldab järjestikku sooritada mitut käsku, nt
cat KOD_* | more avab selles kataloogis kõik failid, mille alguses on KOD_ ning laseb neid ekraanil lehitseda
cat KOD_* | less
"Toru" on väga vajalik edaspidi pikemate otsiridade vms koostamisel.

Kui soovid teada, mitu sõna on avatud failis (st standardväljundisse = ekraanile suunatud failis), saab seda teada käsuga wc:
wc
word count, loeb ridu, sõnu ja sümboleid

Näide:
Et teada saada, kui palju materjali on idamurde alal Kodavere tekstides, liigun kõigepealt idamurde tekste sisaldavasse kataloogi, avan Kodavere failid ning seejärel loen wc-ga sõnad üle.

cd murdekorpus/Idamurre
cat KOD* | wc

Tulemuseks saan kolm arvu, esimene neist näitab ridu, teine sõnade arvu, kolmas sümbolite (tähemärkide) arvu.

```
[108] liina_l@adalberg:~/murdekorpus/Idamurre> cat KOD* | wc
      379    8579   53270
```

Failide salvestamine

Kui faili on kuidagi modifitseeritud, võib selle suunata standatdväljundi (ekraani) asemel ka uude faili. Näiteks on võimalik kõik Kodavere failid korraga avada ning suunata nad ühte faili näiteks mõnes muus kataloogis. Suunamiseks kasutatakse noolt, failinimi ja asukoht tuleb ette anda.

Näide

cd murdekorpus/Idamurre
cat KOD* > /home/murakas/liina_l/kodavere.txt

Nüüd, liikudes vastavasse kataloogi, saate seda faili sirvida või editeerida picot või nanot kasutades.

Masinast väljumine

Sisesta käsureale exit

Failidest otsimine

Failidest märgijada otsimiseks on grep-käsk. See on võrreldav kirjakeele korpuse nn otsiauguga: siin on võimalik koostada samasuguseid päringuskeeme nagu otsimootoriski, ainult tuleb jälgida õiget sisendit ja väljundit. Siin kehtivad ka samad erisümbolid, mis kirjakeele korpuse otsiaugus.

grep
otsib välja kõik vastavat stringi (sõna, sõnaosa) sisaldavad READ

grep 'maja' otsib välja kõik read, kus esineb sõna *maja*
grep - c 'maja' esitab ridade arvu, kus sõna *maja* esineb
grep -v 'maja' otsib välja need read, kus ei esine sõna *maja*
grep -2 'maja' jätab järjendit *maja* sisaldava rea ette ja järele veel kaks rida

Näide
Murdekorpuses on keelejuhi teksti alguses alati märgen < who=KJ>. Kui keelejuhte on mitu, siis vastavalt KJ ja KJ2. Et otsida korpusest AINULT keelejuhi tekstis olevaid sõnu (st keelejuhi kõnevoore), maksab enne täpsustavaid otsinguid välja otsida need read, kus keelejuht räägib, seejärel sealtsamast edasi need read, kus on otsitav tekstilõik, näiteks sõna *vot*:
cat KOD* | grep 'who=KJ' | grep ' vot '

Taolise pärimise järel on ainus probleem, et vastuseks on read, ühe rea pikkus võrdub kogu kõnevooru pikkusega. Selleks, et ridu natuke lühemaks saada, tuleks need tükeldada. Sellest veidi hiljem.

Vihje:
"Toru" puhul tuleb alati arvestada sellega, et ühe käsuga sooritatav väljund on ühtlasi sisendmaterjal järgmise käsu täitmiseks. Seepärast on käskude sooritamise järjekord väga oluline. Kui te täpselt ei tea, mida üks või teine ksäk teeb (st missugune väljund sealt tuleb, mida edasi töödelda kavatsete), tasub töö kõiki etappe üksteise järel läbi katsetada. Sellega hoiate tegevust oma kontrolli all ning leiate hõlpsamalt üles vead, mis võivad tekkida. Eelmises näites võiks eri etappe ükshaaval katsetada nii:

cat KOD* | more siit näed, milline on KOD tekstid, mida edasi töötlema hakkad
cat KOD* | grep 'who=KJ' | more näed, et grep-käsk toimib ja milline on väljund
cat KOD* | grep 'who=KJ' | grep ' vot ' ka siia võib lõppu panna | more , siis on mugavam sirvida.
NB! more-käsu katkestamiseks Cntr+c.

Erisümbolid
töötavad grep- ja sed-käsuga

.	üks suvaline märk
.*	mistahes sümbol null kuni lõpmatu arv kordi
x*	x esineb null kuni lõpmatu arv kordi
[a-z]	inglise tähestiku väiketähed
[A-Z]	inglise tähestiku suurtähed
[0-9]	kõik numbrid
[^a]	kõik märgid, välja arvatud *a*
[^a-zA-Z]	mistahes sümbol, v.a tähestiku tähed
x{2,5}	otsib märgijadasid, kus x-i esineb minimaalselt 2, maksimaalselt 5 korda järjest
|	sooritab kaks otsingut, mis eraldatud |-ga
x{2}	otsib märgijadasid, milles x-i esineb vähemalt 2 korda järjest (maksimum pole piiratud)
x+	otsib märgijadasid, milles x esineb vähemalt 1 korra (maksimum pole piiratud)
\?	kaldkriips tühistab järgneva märgi erisümbolitähenduse
^	rea algus
$	rea lõpp

tr
asendab sümboleid
tr 'a' 'A' asendab kõik väikesed a-d suurte A-dega
tr 'abc' 'efg' asendab kõik *a*-d *e*-dega, kõik *b*-d *f*-idega ja kõik *c*-d *g*-dega
tr '[A-Z]' '[a-z]' asendab kõik suurtähed väiketähtedega
tr -d 'a' kustutab tekstist kõik *a*-d
tr -d '0-9' kustutab kõik numbrid
tr ' ' '\012' asendab kõik tühikud reavahetusega
tr -s '\012' kustutab tekstis kõik korduvad reavahetused, s.t tühjad read

Näide:
Oletame, et on vaja otsida Võru murde tekstidest välja kõik sõnad (mitte read!), mis sisaldavad q-d (larüngaalklusiili). Selleks tuleks eelnevalt 1) otsida sobivad tekstid; 2) otsida neist välja ainult keelejuhi kõnevoorud; 3) asetada iga sõna eraldi reale, et saaks otsida ainult vajalikke sõnu, mitte ridu (see tähendab, et asendan tühikud reavahetusega); 4) otsida välja q-d sisaldavad read.
Saame päringurea (otsin praegu välja kõik Võru tekstid, selleks liigun Võru murde kataloogi):
cd Vorumurre
cat *.txt | grep 'who=KJ' | tr ' ' '\012' | grep 'q' | more
Tulemust vaadates näeme, et vastuste hulgas on ka sõnaühendeid, mille osiste vahel on kokkuhäälduse märk =. Parem oleks ka need asendada reavahetusega:
cat *.txt | grep 'who=KJ' | tr ' ' '\012' | tr '=' '\012' | grep 'q' | more

sed

sobib kasutada siis, kui asendada on vaja märgijada, mitte üksiksümboleid

asendab ühe stringi (sümboliterea, nt sõna) teisega

sed 's/maja/auto/g' asendab kõik *maja*-sõnad sõnaga *auto* (s ja g märgivad, et tegemist on regulaarse asendusega)

sed 's/ä/ä/g' asendab kõik html-i koodis ä-d tavaliste ä-dega

sed 's/(...)/ /g' asendab (...) tühikuga

Näide

Murdekorpusest keelejuhi tekstist sõnade otsimisel on voorud üldjuhul tülikalt pikad. Selleks, et neid lühendada, on mitmeid võimalusi. Küllaltki loogiline koht nende lühendamiseks on pikkade pauside koha pealt: pikad pausid on märgitud (...). Kuna tegemist on märgijadaga (mitte üksiksümboliga), ei saa seda otse reavahetuse koodiga asendada. Mõistlik on enne (...) asendada ühe sümboliga, mida korpuses muidu pole kasutatud, seejärel see üksiksümbol tr-käsuga asendada reavahetuse koodiga. . Näiteks võib (...) asendada X-iga (sed-käsuga), seejärel saab juba tr-käsuga sellesama X-i asendada reavahetusega.

cat *.txt | grep 'who=KJ' | sed 's/(...)/X/g' | tr 'X' '\012'

Lisaks ridade tükeldamisele on mõnikord vaja lahti saada ka veel märgendusest: vooru alguse ja lõpu märkidest, kommentaaridest.
Iga vooru algul on märgitud vooru alguse märk u ja kõneleja: <u who=KJ>
Iga vooru lõpus on vooru lõpu märk </u>

Kommentaaride alguses ja lõpus on <com> kommentaar </com>, seega on kommnetaari märkide vahel litereerija sõnu, mis võivad segi minna keelejuhi või küsilteja päris tekstiga.
Seega oleks vaja vabaneda 1) kõigepealt kommetaaridest; 2) seejärel veel järele jäänud märgenditest.

1) Kommentaaridest vabanemiseks asendame mittemillegiga (või tühikuga, maitse asi) sed-käsu abil lõigu, mis algab <com>, selle vahel on ükskõik mis, v.a < (et ei ületatakse kommentaaride lõpumärki) ükskõik kui palju, seejärel on </com>. Kuna märgendis </com> / on ise sed-käsu sees väljade eristaja, on vaja tema ette panna tagurpidi kaldkriips, et talt ära võtta erisümboli tähendust:
sed 's/<com>[^<]*<\/com>//g'

2) muudest märgenditest vabanemine on juba lihtne, sest märgendite vahelt ei ole vaja kustutada. Asendame mittemillegagi (või tühikuga) järjendi < , seejärel ükskõik mis, mis poleks märgendi lõpusümbol (ja seda ükskõik kui palju), seejärel märgendi lõpusümbol >
sed 's/<[^>]*>//g'

Paneme nüüd kokku käsujärjendi, milles kustutatakse kommentaarid ning hakitakse tekst lühemaks. Suuname ta praegu eraldi faili: murded.txt

Kogu käsk: cat *.txt | grep 'who=KJ' | sed 's/<com>[^<]*<\/com>//g' | sed 's/<[^>]*>//g' | sed 's/(...)/X/g' | tr 'X' '\012' > murded.txt

Faili suunamise asemel võib ka sirvimiseks standardväljundisse suunata (lehitsemisega):
cat *.txt | grep 'who=KJ' | sed 's/<com>[^<]*<\/com>//g' | sed 's/<[^>]*>//g' | sed 's/(...)/X/g' | tr 'X' '\012' | more

Nüüd on teksti niipalju puhastatud, et võib juba midagi asjalikku otsima hakata ☺

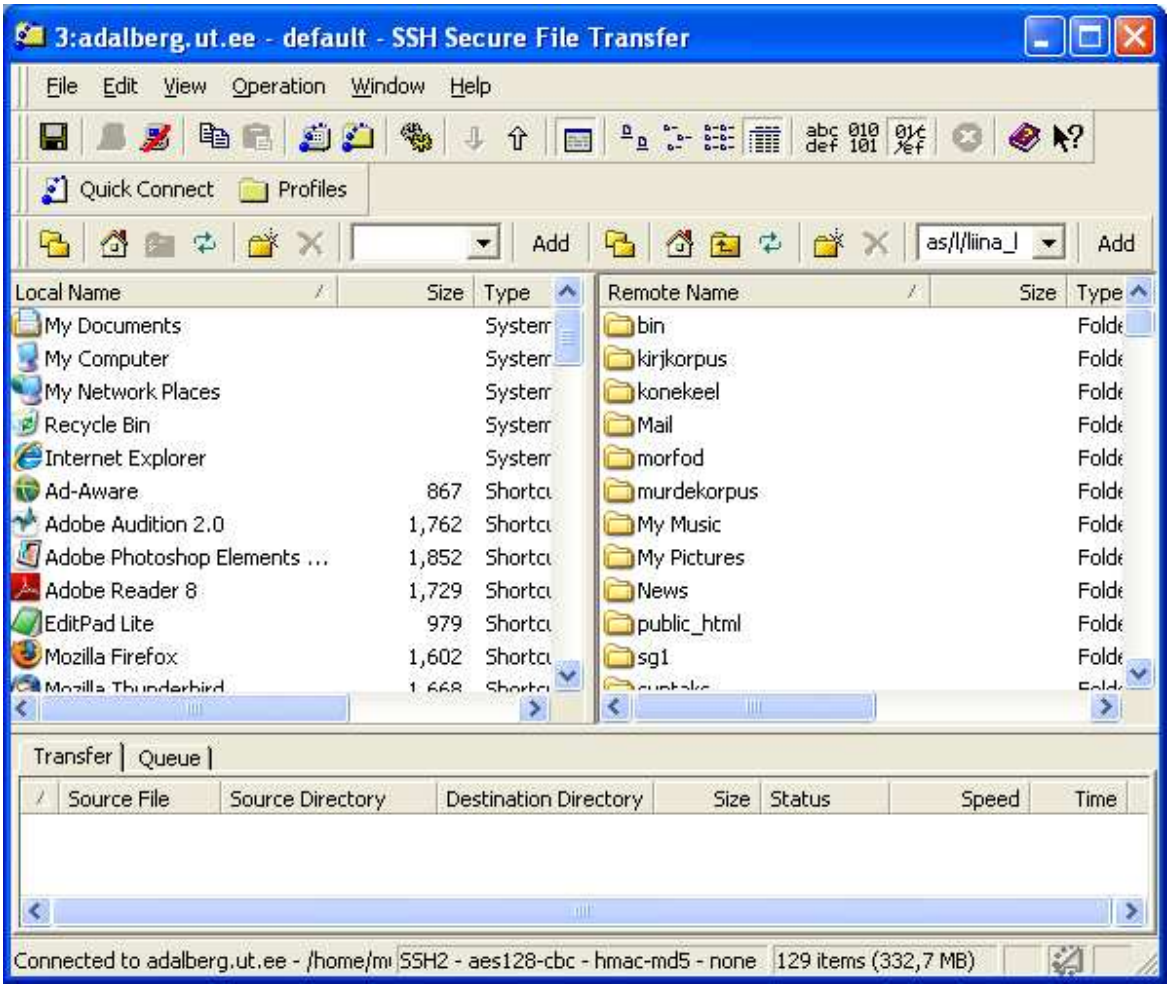
Oma materjali käsitlemine Unixis

Eelnevad näited puudutasid murdekorpuse tekste, mis paiknesid nn õppekorpuses liina_l kodukataloogis. Unixis saab töödelda aga ka muid tekste, näiteks kirjakeele korpuse tekste. Järgnevas osas on kõigepealt juhised, kuidas oma materjali viia oma kodukataloogi ülikooli arvutivõrgus. Oma materjaliks võtame õppimise jaoks kirjakeele korpuse tekstid.

Teema põhiosa käsitleb sagedussõnastikke. Unixis on kerge vaevaga võimalik tekitada kõikvõimalikke sagedussõnastikke, mis sisaldavad keeleuurijale vajalikku informatsiooni. Sagedussõnastike tegemine on iseenesest väga hõlpus, selleks on vaja ainult mõnda käsku kombineerida.

Failide viimine Unixisse

Failide tõstmiseks Unix-keskkonda kasutage taas programmi SSH Secure Shell Client, ent valige rippmenüüst New File Transfer. Seejäre vajutage nupule Qick Connect ning sisestage dialoogiaknasse taas masina nimi, kuhu tahete end sisse logida (adalberg.ut.ee) ja oma kasutajanimi, seejärel vajutage Connect. Seejärel sisestage uude dialoogiaknasse oma parool. Kui olete end edukalt sisse loginud, saate kahest poolest koosneva ekraanipildi. Vasakul paikneb selle arvuti sisu, milles parasjagu olete (teie lauaarvuti, läpakas vms), paremal pool aga teie kodukataloogi sisu ülikooli arvutivõrgus.



Vajalike failide tõstmiseks oma masinast adalbergi või vastupidi valige sobiv kataloog oma nasinast (vasakul pool) ja sobiv kataloog adalbergis (parem pool), vajadusel tekitage adalbergi uus kataloog. Failide tõstmiseks lohistage nad hiirega ühelt poolt teisele.

Kirjakeele korpuse materjalide tõstmine adalbergi

Et kasutada kirjakeele korpuse materjale Unixis, on üks võimalus need kõigepealt alla laadida oma arvutisse, seejärel lahti pakkida (nad on pakitud zip-failiks) ja alles seejärel tõsta tekstid Unixisse. Kirjakeele korpuse materjalid leiate korpuse sisututvustuse alt, näiteks 1930ndate tekstid paiknevad siin:
http://www.cl.ut.ee/korpused/baaskorpus/1930/

Laadime siit ilukirjandustekstid oma arvutisse ja pakime lahti, selleks:

1) klõpsake hiirega zip-failil (Ilukirjandustekstid), seejärel küsitakse, kas tahate faili avada või salvestada. Valige salvestamine, otsige või tehke sobiv kataloog (nt ilu1930).

2) Pakkige failid lahti: Windowsis klõpsake parempoolse hiirenupuga failinimel ning valige rippmenüüst Extract all.
3) Saate ühe suure faili, millel ei ole failinimelaiendit, seepärast ei oska windows ise valida, mis programmiga seda avada. Kui tahate faili Windowsis vaadata, vaadake seda näiteks Wordpadiga.
4) Tõstke fail Unixisse (juhend eespool).

Nüüd on fail valmis kasutamiseks. Vahetame failivahetusakna taas terminaaliakna vastu (menüüst Window valige New terminal), olete samas masinas.

Kõigepealt võiksite faili lihtsalt sirvida, et meelde tuletada, kuidas see on organiseeritud: cat failinimi | more

Näete, et iga rea alguses on kood, sellele järgneb neli tühikut. Täpitähed on html-kujul.

Faili saab kasutada ka samasuguste otsingute tegemiseks, nagu tegime veebipõhise kirjakeele korpuse otsimootriga. Oluline on aga kogu aeg meeles pidada sisendit ja väljundit. Järgmises näites on lihtne näide selle kohta, kuidas tekstist otsida grep-käsuga.

Harjuta: täiendlausete leidmine

Täiend- ehk relatiivlaused iseloomustavad mingit kindlat nimisõna:
Mees, kes jooksis üle tänava, oli minu kauge sugulane.

Relatiivlaused algavad eesti keeles tüüpiliselt küsiv-siduvate sõnadega *kes/mis*, mis võivad olla käänatud (üldjuhul ainult ainsuses, *mehed, kelledega me läksime....* on väga ebatavaline.)
Piiramegi otsingu praegu nendele sõnadele ja vormidele. (Teoreetiliselt on ka muid võimalusi, nt *kus*.)
Seega lähtume otsingul järgmisest vormistikust:

kes	mis
kelle	mille
keda	mida
kellesse	millesse jne

Seega otsime märgijadasid ke[sdl] ja [mi[sdl], nende ühisosa oleks [km][ei][sdl].
Arvestame veel, et relatiivlause järgneb kirjakeele reeglite järgi komale. Seega:
1) avame faili: cat 30_ilu_ttxt
2) otsime järjestust , ke[sdl] või , mi[sdl]: grep ', [km][ei][sdl]'
3) sirvime faili more-iga või suuname uude faili

Kogu käsk: cat 30_ilu_ttxt | grep ', [km][ei][sdl]' | more

11. Sagedussõnastikud

Kursuse viimane teema puudutab lihtsate sagedussõnastike tegemist Unixis/Linuxis. Sagedussõnastikke kasutatakse keeleteaduses küllalt palju ja need sisaldavad ohtralt lingvistile vajalikku informatsiooni. Sagedussõnastikke on ka välja antud, näiteks kirjakeele korpuse põhjal on Heiki-Jaan Kaalepi ja Kadri Muischneki koostatud sagedussõnastik, selle veebiversiooni näete siit:

http://www.cl.ut.ee/ressursid/sagedused/

Lihtsat sagedussõnastikku on kasutatud ka mitmetes muudes uurimustes. Kui nipp on käes, võib sagedussõnastiku teha igast tekstist, mis ette juhtub!

Sagedussõnastiku materjali ettevalmistamine

Et tekstidest hakata sagedussõnastikke tegema, oleks kõigepealt vaja vabaneda koodist rea algul – vastasel juhul arvestame sagedussõnastikes ka lausekoode kui sõnu. Koodist on kõige hõlpsam vabaneda cut-käsu abil.

Sagedussõnastiku tegemiseks oleks vaja veel asendada kõik suurtähed väikestega (sest sõnavormid *Kõik* ja *kõik* on ju sama sõna vormid, neid on vaja koos arvestada); samuti on tülikad täpitähed. Et vabaneda html-kujul täpitähtedest ning asendada need tavalistega, võiks (vähemalt sagedasemad) sed-käsuga asendada.

cut
lõikab reast välja etteantud välja. Väli tuleb ise defineerida lipukesega -d
cut -d " " jutumärkide vahele tuleb sümbol, mis välju piiritleb, antud juhul tühik
-f1 number märgib välja numbrit, anutd juhul 1. väli (seega enne tühikut, kui tühik piiritleb välju; sisuliselt on otsitakse selles näites välja rea esimene sõna)
cut –d" " –f1-3 lõikab välja esimesed kolm välja, väljad on defineeritud tühikutega (=sõnad)
cut –c1-2 lõikab reast välja ja suunab standardväljundisse esimesed kaks sümbolit

Näide: faili ettevalmistamine sagedussõnastiku tegemiseks
Koodist vabanemiseks avame faili, seejärel defineerime tühiku väljade eristajana cut-käsus ning lõikame välja 5. väljast alates kõik:

cat 30_ilu_ttxt | cut -d" " -f5- | more

Suurtähtede asendamine väiketähtedega: tr '[A-Z]' '[a-z]'
cat 30_ilu_ttxt | cut -d" " -f5- | tr '[A-Z]' '[a-z]' | more

Täpitähtedest vabanemine (NB! kuna suured tähed on juba asendatud väiketähtedega, pole ka suurte Ä-dega jne vaja arvestada):
cat 30_ilu_ttxt | cut -d" " -f5- | tr '[A-Z]' '[a-z]' | sed 's/ä/ä/g' | sed 's/ö/ö/g' | sed 's/ü/ü/g' | sed 's/õ/õ/g' | more

Täpitähtede sisestamisel võib tekkida probleeme.
Kuna seda rida on tülikas pidevalt korrata, on mõistlik tekitada endale fail, kus need eeltööd on tehtud, ning hiljem teha otsinguid /töödelda edasi seda faili:
cat 30_ilu_ttxt | cut -d" " -f5- | tr '[A-Z]' '[a-z]' | sed 's/ä/ä/g' | sed 's/ö/ö/g' | sed 's/ü/ü/g' | sed 's/õ/õ/g' >1930ilu

Edaspidi võtta sisendiks see fail, näiteks cat 1930ilu | more

Sagedussõnastiku koostamine

Sagedussõnastike tegemiseks on vaja, et teksti oleks eeltöödeldud, st eemaldatud oleks kood rea algusest, suurtähed oleks asendatud väiketähtedega, samuti oleks mõistlik kustutada kirjavahemärke. Kuna eelmises näites koodi kustutamist ja suurtähtede asendamist näitasime, ei hakka seda siin kordama. Kui olete faili juba eeltöödelnud ja selle ka salvestanud, kasutage sisendina seda eeltöödeldud faili.

Sagedussõnastiku jaoks oleks hea kustutada ka kirjavahemärgid. Selle vajalikkus sõltub peamiselt konkreetsest korpusest: kui kirjavahemärgid on juba muust tekstist tühikuga eraldatud, ei hakka need sagedussõnastiku tulemusi mõjutama, kirjavahemärk loetakse siis nagu omaette sõna ja reeglina paiknevad sagedussõnastiku sagedasemas otsas. Sealt võib nad lihtsalt pärast välja visata.

Kui kirjavahemärgid ei ole muust tekstist tühikutega eraldatud, on kasulik nad eelnevalt kustutada; vastasel juhul loetakse näiteks sõnavormid *mees, mees. mees! mees? mees* kõik erinevateks sõnadeks sagedussõnastikus.

Kirjavahemärkide kustutamiseks kasutame käsku tr, lipukese–d taha ülakomade vahele lisame kõik märgid, mis oleks vaja kustutada (kontrolli kindlasti, milliseid kirjavahemärke selles tekstis on kasutatud):
tr –d '.,!?<>"-'

Seejärel oleks vaja iga sõna asetada eraldi reale. Selleks asendasime tühikud reavahetusega:
tr ' ' '\012'

Seejärel 1) sorteerime read; 2) kustutame korduvad read (nii, et jääb alles arv, mitu korda rida=sõna esines); 3) võime sorteerida saadu tagurpidises järjekorras (kuna sagedusinfo paigutatakse rea ette, reastatakse esinemissagedus normaaljuhul vähimast kordade arvust suurimani; kui aga reastame tagurpidi järjekorras, siis suurimast vähimani).

sort
sorteerib read tähestiku järjekorras
sort -f ei tee vahet väike- ja suurtähtedel
sort -r sorteerib vastupidises järjekorras

uniq
kustutab korduvad read
uniq -c lisab rea ette arvu, mis näitab, mitu korda seda rida esines

Näide: sagedussõnastiku tegemine

1) Puhastame teksti
cat 30_ilu_ttxt | cut -d" " -f5- | tr '[A-Z]' '[a-z]' | tr -d ',.?!<>:'
Võime kasutada ka juba eelpuhastatud faili sisendina.
Kui täpitähed html-kujul häirivad, võite ka need asendada (vt eespoolt).

2) asetame iga sõna eraldi reale: tr ' ' '\012'
kogu käsk:
cat 30_ilu_ttxt | cut -d" " -f5- | tr '[A-Z]' '[a-z]' | tr -d ',.?!<>:' | tr ' ' '\012'

3) sorteerime, kustutame korduvad read: sort | uniq –c
kogu käsk:
cat 30_ilu_ttxt | cut -d" " -f5- | tr '[A-Z]' '[a-z]' | tr -d ',.?!<>:' | tr ' ' '\012' | sort | uniq –c

4) sorteerime veelkord tagurpidi järjestuses: sort -r

kogu käsk:
cat 30_ilu_ttxt | cut -d" " -f5- | tr '[A-Z]' '[a-z]' | tr -d ',.?!<>:' | tr ' ' '\012' | sort | uniq –c | sort -r

Näide: millega algab küsilause kõige sagedamini?

Kui tahame teada, millega algab küsilause kõige sagedamini, tuleb kõigepealt mõelda välja, kuidas leida küsilaused, seejärel lõigata välja nende lausete esimesed sõnad ning teha neist sagedussõnastik.

1) Valime faili: cat 30_ilu_ttxt
2) otsime välja küsilaused: grep '?'
3) lõikame välja lause esimese sõna. Selleks defineerime tühiku kui väljade eristaja, ja nagu varem rehkendasime, on sel juhul lause esimene sõna 5. väli. Kuna lause lõpuni pole vaja, siis näeb see käsk välja nii: cut –d' ' –f5
   Kuna lõikasime välja vaid ühe sõna lausest, pole vaja enam reavahetusi tühikutega asendada.
4) Seekord pole vaja tingimata ka suuri tähti väikestega asendada, sest lause algul peaksid kõik sõnad algama suure tähega. Seega võime jätkata sorteerimise ja sageduse järgi reastamisega: sort | uniq –c | sort -r

Kogu käsk: cat 30_ilu_ttxt | grep '?' | cut –d' ' –f5 | sort | uniq –c | sort –r

Näeme, et 1930ndate tekstist otsides häirib otsingut « - see peaks olema jutumärk html-is. Enne sagedussõnastiku tegemist võiks selle osa sed-käsuga kustutada, st asendada mittemillegagi. Samuti võiks eelnevalt kustutada jutumärgid.

Kogu käsk: cat 30_ilu_ttxt | grep '?' | sed 's/«//g' | tr –d '"' | cut –d' ' –f5 | sort | uniq –c | sort –r
Tulemusi sirvides näeme, et palju on hulgas sõnu, mis pole küsisõnad. Need on tingitud lausetest, kus näiteks teine või kolmas osalause on küsilause.

Näide: millise kaashäälikuühendiga algavad sõnad kõige sagedamini?

Sagedussõnastikke ei pea tegema ainult sõnadest, vaid võib teha ka sõnaosadest vms. See näide puudutab sõnaalgulisi kaashäälikuühendeid. Nagu teame, on kaashäälikuühend sõna alguseses eesti keeles seotud üldjuhul laensõnadega, mitte omasõnadega, ning nende hulk on piiratud. Missugused on aga kõige tavalisemad?

Selleks, et kaashäälikuühenditest sagedusnimestikku teha, tuleb arvestada, et 1) igasuguste asjade väljaotsimine grep-käsuga toimub reakaupa. Selleks on vaja kõik sõnad tõsta eraldi reale, samuti suurtähed asendada väiketähtedega; 2) jätta välja kõik kahetähelised lühendid: *st, jt* jne. Selleks võiks mingil kompel defineerida sõna pikkuse, ilmselt peaks sõna olema pikem kui kolm tähte. 3) tuleb välja otsida iga sõna esimesed kaks tähte ning nende hulgast välja valida need, milles mõlemad on konsonandid. Neist saab siis teha sagedussõnastiku. Töö etappide kaupa:

5) Valime faili: cat 30_ilu_ttxt
6) Asendame iga sõna eraldi reale, asendame suurtähed väiketähtedega: tr ' ' '\012' | tr '[A-Z]' '[a-z]' Sellega 1 sõna=1 rida. Kui on veel segavaid märke, kustutame need: tr –d',.?!:"-'
7) otsime välja read, mis sisaldavad vähemalt nelja tähte: grep '....'
8) lõikame välja iga rea esimesed kaks tähte: cut –c1-2
9) otsime nende hulgast välja need, milles nii eismene kui teine oleks kaashäälikud: grep '[bdghjklmnprstv][bdghjklmnprstv] '
10) teeme neist sagedussõnastiku: sort | uniq –c | sort-r

Kogu käsk: cat 30_ilu_ttxt | tr ' ' '\012' | tr '[A-Z]' '[a-z]' | tr –d',.?!:"-' | grep '....' | cut –c1-2 | grep '[bdghjklmnprstv][bdghjklmnprstv]' | sort | uniq –c | sort -r